Module code: M303 **Further Pure Mathematics**

Resources

Contents

1. Resources Page
2. History Reader
3. Accessibility for Maths
4. Book A Figure Descriptions
5. Book B Figure Descriptions
6. Book C Figure Descriptions
7. Book D Figure Descriptions
8. Book E Figure Descriptions
9. Book F Figure Descriptions
10. Good Academic practice for M303

SUP088757

Resources

Downloads ›

Library resources ›

Resource settings

Resources advice

Navigation

Administration

General items

Tutor website

Hidden from students

M303 Guide

Pay particular attention to the section on assessment and the paragraph about the Handbook.

History Reader | 5.4MB PDF document

Useful resources for M303 (external site)

This link goes to a website that has some useful resources for M303. Please note this is an external website and The Open University is not responsible for its content.

M303-19J Student common room

You are welcome to use this room to have online meetings with other students to discuss the module.
For students only.

Informal lecture: introduction to sets | 60.3MB

You may find this helpful for TMA04 and TMA06. We discuss set notation and how to sketch sets in the plane. This is a recording from a previous presentation that, after a lot of work on Hayley's part, was made into an mp4.

Accessibility

An accessibility guide has not been prepared specifically for M303 but ‘Accessibility for mathematics and statistics modules’ outlines some of the general accessibility issues that apply to distance learning of mathematics and statistics.

Accessibility for mathematics and statistics modules

Books A to F: figure descriptions

Handbook

The M303 Handbook is intended to supplement the module and to provide easy reference to all the important definitions and results from the module, and for use during the examination.

This module covers several areas of pure mathematics and is divided into six books, so the Handbook also has six sections, one for each book, in addition to a ‘Quick reference’ section and some appendices of further useful information.

Each individual book section comprises four chapter summaries with all the definitions and results from the book and (possibly) a short Glossary containing items not covered by the chapter summaries. The glossary entries are quite brief, but they provide a reference to the relevant chapter and section of the book, where you can find more information.

The index is intended to be quite comprehensive, so, to find entries in the Handbook, the advice is to look there first, unless you know you are looking for a result or definition in a specific chapter.

You can take the printed copy of the Handbook into the exam, but please note that only a basic level of annotation will be permitted. ‘Basic’ annotation is defined as follows:

> "The text as printed may be supplemented by handwritten highlights (for example by a highlighter pen or by ringing, underlining or sidelining), and by corrected typographical errors. The addition of comments, marginal notes, notes in the blank spaces at the end of paragraphs and pages or on fly-leaves is not permitted".

Note that only the printed version sent to you can be taken into the exam, not your own printed versions or photocopies.

Handbook | 1.3MB PDF document

Chapters

Individual chapter PDF files are supplied here for your convenience.

Books A to F: chapter PDFs

M208 units

We have provided some M208 units that you may find useful to revise as you study M303.

M208 units

MST 125 unit on proof

MST 125 unit 9: Mathematical language and proof | 4.9MB

Computing guide

Computing guide

A link to the OU generic Computing guide.

Careers and employability

Careers and employability

Mathematics and statistics study website

Undergraduate mathematics and statistics site

To support your M&S studies this site covers:

- CONNECT – to link with the M&S community at the OU and beyond
- DISCOVER – check you are ready for your module, and make a head start
- SKILLS – guidance on TMA and exam preparation, etc specifically for M&S
- PLAN – advice on what to study next
- SUCCEED – information on relevant careers and developing your employability

Feedback

© 2019. All rights reserved. The Open University is incorporated by Royal Charter (RC 000391), an exempt charity in England & Wales and a charity registered in Scotland (SC 038302). The Open University is authorised and regulated by the Financial Conduct Authority in relation to its secondary activity of credit

broking.

The Open University

M303

Further pure mathematics

History Reader

This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from the Student Registration and Enquiry Service, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)845 300 6090; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

To purchase a selection of Open University materials visit www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom for a brochure (tel. +44 (0)1908 858779; fax +44 (0)1908 858787; email ouw-customer-services@open.ac.uk).

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2014.

Copyright © 2014 The Open University

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, transmitted or utilised in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without written permission from the publisher or a licence from the Copyright Licensing Agency Ltd. Details of such licences (for reprographic reproduction) may be obtained from the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS (website www.cla.co.uk).

Open University materials may also be made available in electronic formats for use by students of the University. All rights, including copyright and related rights and database rights, in electronic materials and their contents are owned by or licensed to The Open University, or otherwise used by The Open University as permitted by applicable law.

In using electronic materials and their contents you agree that your use will be solely for the purposes of following an Open University course of study or otherwise as licensed by The Open University or its assigns.

Except as permitted above you undertake not to copy, store in any medium (including electronic storage or use in a website), distribute, transmit or retransmit, broadcast, modify or show in public such electronic materials in whole or in part without the prior written consent of The Open University or in accordance with the Copyright, Designs and Patents Act 1988.

Edited, designed and typeset by The Open University, using the Open University TEX System.

WEB 03766 2

1.1

Contents

1 The origins of structural algebra **4**
1.1 From the solution of equations to Galois Theory 4
1.2 From Galois Theory to finite group theory 19
1.3 Finite fields (Galois, Jordan, Steinitz) 23

2 The origins of topology and metric spaces **26**
2.1 The importance of Fourier series 26
2.2 Potential theory and the Dirichlet principle 38
2.3 Metric spaces – Fréchet 44
2.4 Topological spaces – Hausdorff 45

3 Biographies **48**
3.1 Pythagoras 48
3.2 Euclid 49
3.3 Marin Mersenne 49
3.4 Pierre de Fermat 50
3.5 Leonhard Euler 51
3.6 Joseph-Louis Lagrange 53
3.7 Adrien-Marie Legendre 53
3.8 Joseph Fourier 55
3.9 Carl Friedrich Gauss 56
3.10 Peter Gustav Lejeune Dirichlet 59
3.11 Bernard Riemann 60
3.12 Georg Cantor 60
3.13 Felix Hausdorff 61

References **62**

Acknowledgements **65**

1 The origins of structural algebra

1.1 From the solution of equations to Galois Theory

It can be hard to believe that modern structural algebra (groups, rings, fields, homomorphisms and isomorphisms) grew out of straightforward mathematical problems, and that the algebra we study at school is its legitimate ancestor. But such is the case, and it can be instructive to see why. There are, in fact, few more dramatic stories in all of mathematics than the way that attempts to solve polynomial equations turned into the study of groups and fields, as we will now describe.

Cubic and quartic equations

Mathematicians in ancient Mesopotamia some 4000 years ago knew how to solve problems we would write today as quadratic equations. Here is one of their surviving problems. (The form of the numbers has been modernised.)

From Robson (2007, p. 104).

> I took away my square-side from inside the area and it was 870. You put down 1, the projection. You break off half of 1. You combine 0.5 and 0.5. You add 0.25 to 870. 870.25 squares 29.5. You add 0.5 which you combined to 29.5 so that the square-side is 30.

The first line tells us that the problem is to find the side of a square when the area minus the side is 870. The modern version of this would be the equation $x^2 - x = 870$, which we are to solve for x. We can follow what is going on by looking at what we do to solve $ax^2 - bx - c = 0$. The method is to take 1 (b, the coefficient of x, so to speak), halve it and square it to give 0.25 (or $(b/2)^2$), add it to the 870 (which is ac in our equation), take the square root ($\sqrt{(b/2)^2 - ac}$), add 0.5 (which is $b/2$) and thus obtain the answer (the one we learn today as $\dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$). Notice that in the Mesopotamian problem c is negative. The Mesopotamian scribe naturally considered only the positive square root.

A perfectly general theory of quadratic equations having at least one positive real root was written down in the ninth century CE by the Islamic mathematician Al-Khwarizmi, but there the matter rested until, by the early sixteenth century, it had become something of a scandal that no progress had been made on solving cubic equations in terms of an algebraic formula of some kind. There was then a flurry of activity, spurred on by a very public rivalry between the leading contestants (as success at problem solving could lead to a lucrative contract as a scientific advisor at court), in which honours were finally divided between Niccolò Tartaglia and Gerolamo Cardano (Figure 1.1).

See the Foreword in Cardano (1543).

In those days mathematicians distinguished sharply between equations such as $x^3 + ax = b$ and $x^3 + b = ax$, because they required that the coefficients a and b be positive in each case. A mathematician in Bologna called Scipione del Ferro found a way of solving some cubic equations

(those of the form $x^3 + ax = b$, which have one positive real root – the others are complex and eluded del Ferro) and he confided it to a pupil, Mario Fior. Fior in turn challenged Tartaglia to a problem-solving competition, thinking that with his secret method he was bound to win, but the night before the competition Tartaglia found the way to solve Fior's cubics, and won. The fame he acquired brought him to the attention of Cardano, and Tartaglia confided his methods to Cardano under conditions of strict secrecy in exchange for an introduction to the local prince.

Figure 1.1 Gerolamo Cardano (1501–1576)

Cardano then found out how to solve all types of cubic equation, and his assistant, Ludovico Ferrari, showed how to solve equations of degree 4 (quartic equations). They chafed for some time under Cardano's promise because they had so much new to say, until they heard about Scipione del Ferro and Fior; they then decided that Tartaglia had not, after all, possessed a secret method, so it was permissible to publish what they had discovered. Tartaglia was predictably furious and a long row began between him and Ferrari. But the book that Cardano published, the *Ars Magna* (*Great Art*) of 1543, moved algebra out of the limited domain of equation-solving it had inhabited for centuries.

The general cubic equation

$$ax^3 + bx^2 + cx + d = 0$$

can always be written in a simpler form. We divide through by a to get an equation of the form

$$x^3 + b'x^2 + c'x + d' = 0.$$

If we replace x with $x - \frac{b'}{3}$ we get

$$(x - \tfrac{b'}{3})^3 + b'(x - \tfrac{b'}{3})^2 + c'(x - \tfrac{b'}{3}) + d' = 0.$$

The coefficient of x^2 is $-3\frac{b'}{3} + b' = 0$, so the equation now has this form:

$$x^3 + px + q = 0.$$

It is the equation in this form that we solve; it is then merely a tedious exercise to work back to the solution of the original equation.

Tartaglia, and Cardano after him, exploited the similarity in form between the above equation and

$$(u + v)^3 - 3uv(u + v) - (u^3 + v^3) = 0.$$

This is an algebraic identity, true for all values of u and v. So if we set $p = -3uv$ and $q = -(u^3 + v^3)$, it follows that x is $u + v$. It remains to solve this pair of simultaneous equations for the unknowns u and v, and the solutions turn out to be

$$u = \sqrt[3]{-\tfrac{q}{2} + \sqrt{(\tfrac{q}{2})^2 + (\tfrac{p}{3})^3}},$$

$$v = \sqrt[3]{-\tfrac{q}{2} - \sqrt{(\tfrac{q}{2})^2 + (\tfrac{p}{3})^3}}.$$

Notice that in each case we have three cube roots to consider: $\omega^k u$ and $\omega^m v,\ k, m = 0, 1, 2$, where ω is a primitive cube root of unity ($\omega \neq 1$, $\omega^3 = 1$). A little algebra shows that the three solutions of the cubic are

$$u + v,\ \omega u + \omega^2 v,\ \omega^2 u + \omega v.$$

Important though this achievement is, the formula is often of little value. Consider, as Cardano did, this cubic:

$$x^3 - 7x + 6 = 0.$$

It has three real roots $x = 1, 2, -3$, but in this case $(\frac{q}{2})^2 + (\frac{p}{3})^3 = -\frac{100}{27}$ is negative, so the formula involves complex numbers. But for cubic equations with only one real root and two complex roots, the formula gives the real root without involving complex numbers, so the occurrence of complex numbers is counter-intuitive. This is why cubic equations mark the point at which imaginary numbers forced themselves into mathematics. They did not come in with quadratic equations but with cubics, where the value of the roots is completely obscured unless an algebra is developed to handle these imaginary numbers.

Cardano's assistant Ludovico Ferrari seems to have been so violent at times that Cardano wrote in his biography of Ferrari that he was frightened of him. However, he turned out to be the mathematician who discovered how to solve the quartic equation, and he always acknowledged a lifelong debt to Cardano. The trick that we used to remove the quadratic term from a cubic equation can also be used to remove the cubic term from an equation of degree 4, so the equation can always be put in the form

See Cardano (1663, pp. 568–9) and Morley (1854, p. 266).

$$x^4 + px^2 + qx + r = 0.$$

Ferrari had the idea of putting x^4 on one side and all the other terms on the other side, then completing the square for the terms on the right-hand side. So he wrote

$$x^4 = -px^2 - qx - r,$$

and added $2zx^2 + z^2$ to each side. This gave him

$$x^4 + 2zx^2 + z^2 = (2z - p)x^2 - qx + (z^2 - r),$$

and he now looked for the value of z that makes the right-hand side a perfect square. This is found, by the $b^2 - 4ac$ test, to be a root of

$$q^2 = 4(2z - p)(z^2 - r).$$

This is a cubic in z and so, by Cardano's earlier work, it can be solved. Taking any value of z that is a solution of this equation, we can solve the corresponding equation for x^2 and take square roots to get the required values of x.

Quintic equations – Lagrange's work

The solution of quadratic, cubic and quartic equations involves the usual operations of arithmetic (addition, subtraction, multiplication and division) and the taking of square or cube roots. For this reason the solution method is called 'solving by radicals', from the Latin word *radix* for a root.

But progress halted with the quartic equation, and attempts by several mathematicians to solve the general quintic equation by radicals all failed. On the one hand, it was increasingly clear to mathematicians that every polynomial equation has as many roots as its degree indicates (this is the content of the fundamental theorem of algebra, which was only proved in the early nineteenth century but was long believed). On the other hand, no one could find an explicit formula involving nothing but addition, subtraction, multiplication, division and root extraction that exhibited the roots of a polynomial equation of degree 5 in terms of the coefficients of the equation.

The first person to take a fresh look at the problem was Joseph-Louis Lagrange. His plan was to investigate why cubic and quartic equations can be solved, and then either to see if the quintic can be solved for the same sort of reason, or to conclude that perhaps the quintic cannot in general be solved by radicals.

See Lagrange (1770).

The spur to Lagrange's thinking was this. A cubic equation $x^3 + px + q = 0$ has three roots; let us call them u, v, w. We have

$$u + v + w = 0,\ uv + vw + wu = p,\ uvw = -q.$$

Now, the expressions $u + v + w$, $uv + vw + wu$ and uvw are symmetric; they do not change their values under permutations of u, v and w. In fact, any symmetrical expression in u, v and w can be written in terms of these three. But the formula for the roots involves square roots and cube roots, and square roots take only two values and cube roots only three. The converse is also true. For example, an expression in u, v and w that takes only two values should be the root of a quadratic equation whose coefficients are rational functions of p and q.

Suppose σ_1 is such an expression, and under the six permutations of u, v and w its other value is σ_2. Consider the equation

$$x^2 - (\sigma_1 + \sigma_2)x + \sigma_1\sigma_2 = 0.$$

A permutation either leaves σ_1 and σ_2 alone or exchanges their values, so any permutation of u, v and w leaves $\sigma_1 + \sigma_2$ and $\sigma_1\sigma_2$ alone. But that means that $\sigma_1 + \sigma_2$ and $\sigma_1\sigma_2$ are expressible in terms of p and q.

A very similar argument deals with expressions in u, v and w that take only three values. So Lagrange speculated that the cubic equation is solvable because there are expressions in u, v and w that take only two or three values.

In this case, the key expression is $\sigma = (u - v)(v - w)(w - u)$, which takes just two values under the six permutations: σ and $-\sigma$. This means that σ^2 is symmetric, so it is an expression in p and q. Since p is of degree 2 in u and v, q is of degree 3 and σ^2 is of degree 6, this suggests trying expressions in p^3 and q^2, which recalls the formula for the roots. A little algebra reveals that

$$\sigma^2 = -4p^3 - 27q^2,$$

from which we deduce that

$$\pm\frac{\sigma}{6\sqrt{-3}} = \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}.$$

So, to explain why the cubic equation is solvable by radicals, Lagrange argued that it is the existence of an expression in the roots that takes only two values that accounts for the existence of the square root in the formula, and then the solution is completed by nothing more elaborate than taking a pure cube root.

We can now sketch Lagrange's very lengthy analysis of the quartic equation

$$X^4 + aX^3 + bX^2 + cX + d = 0.$$

Let the roots of this equation be x_1, x_2, x_3 and x_4. From Ferrari's work, it is clear that the solution involves solving a cubic equation and then taking square roots of expressions that involve the roots of that cubic equation. From his analysis of the general cubic equation, Lagrange knew that the fact that the quartic is solvable by radicals is probably due to the existence of expressions in the roots that take fewer than 24 values as the roots are permuted, and indeed take only 2 or 3 values, which would mean they were roots of quadratic or cubic equations.

Compare Book B, Chapter 5, Exercise 5.4.

Lagrange looked at all the known methods for solving the quartic equation, and showed that they all depended on the existence of partially symmetric functions of the roots – for example, the expression $x_1x_2 + x_3x_4$, which takes only three different values, or the expression $x_1 + x_2 - x_3 - x_4$, which takes only six. Towards the end of a long memoir in 1770 he concluded the following (Lagrange, 1770, p. 304):

> We have not only brought together these different methods and shown their connection and interdependence, but we have also – and this is the main point – given the reason, a priori, why some of them lead to resolvents of degree three and others to resolvents of degree six that can be reduced to degree three. This is due, in general, to the fact that the roots of these resolvents are functions of the quantities x_1, x_2, x_3, x_4, which, like the function $x_1x_2 + x_3x_4$, take on just three values under all the permutations of these four quantities; or which, like the function $x_1 + x_2 - x_3 - x_4$, take on six values that are pairwise equal but have opposite signs; or which take on six values that can be separated into three pairs such that the sum or product of the values in each pair is not changed by the permutations of the quantities x_1, x_2, x_3, x_4 .... The general solution of quartics depends uniquely on the existence of such functions.

Accordingly, when it came to the quintic, Lagrange looked for expressions in five symbols x_1, x_2, x_3, x_4, x_5 that took only three or four values as the xs were permuted. Such expressions would satisfy equations of degree 3 or 4, which he called 'resolvents', and these equations would have coefficients that were explicit combinations of the coefficients of the quintic equation, and so could be solved. However, because Lagrange worked with expressions directly, and did not make a separate study of permutations and their combinations, his analysis was very clumsy and long-winded.

This may have contributed to his eventual opinion of the difficulty of the problem of solving the quintic, which was not that it could not be done but that success was unlikely. He also noted that even the solvability of the quartic depended on a happy accident: the expression $x_1 + x_2 - x_3 - x_4$ that takes six values in fact takes three values and their negatives, so it leads not to a typical equation of degree 6 but one that can be solved by solving a quadratic and a cubic equation.

The best that Lagrange could do was to find an expression in the roots that took only six different values when the roots were permuted. But this led to an equation of degree 6, and there seemed to be no reason why it would be one of those equations that could be solved by solving first a quadratic and then a cubic. If it could not, there was no hope of solving the general equation of degree 5 by tackling an equation of higher degree. He expressed his downbeat conclusions this way (Lagrange, 1770, p. 403):

> Here, if I am not mistaken, are the true principles for the solution of equations and for the best analysis of its conduct. As one can see, everything reduces to a kind of combinatorial calculus by which one finds a priori the results to be expected. It may seem reasonable to apply these principles to equations of the fifth and higher degrees, of which the solution is at present unknown, but such application requires too large a number of examinations and combinations whose success is very much in doubt. Thus, for the time being, we must give up this work. We hope to be able to return to it at another time. At this point we shall be satisfied with having set forth the foundations of a theory which appears to us to be new and general.

In fact, Lagrange never returned to the topic and his next long paper on the topic of solving equations was on the method of finding numerical approximations to their roots, which some historians have suggested is further evidence that he despaired of solving the quintic algebraically. In fact, such an approach is consistent with the belief that even if the equations could be solved by radicals, the solution expressions would be unintelligible, as the example of the cubic equation has already suggested.

Little was then done for a further 30 years, until the Italian mathematician Paolo Ruffini announced that the general quintic equation was not solvable by radicals. He was unlucky in not being sympathetically read, for his argument is flawed at only one point and his conclusion is correct. The first fully correct proof was then given in 1824 by the young Norwegian mathematician Niels Henrik Abel (Figure 1.2). Each tackled the problem of what the succession of a sequence of equations of low degree would look like, which is to say they wrote down expressions that are pure n_1th roots of the coefficients of the quintic equation, then expressions that are pure n_2th roots of the coefficients of the quintic equation and the n_1th roots already obtained, and so on. They considered what properties such expressions must have if they contribute to the solution of the quintic, and each showed, Abel more rigorously than Ruffini, that on symmetry grounds no expressions exist that correspond to the roots of the general quintic.

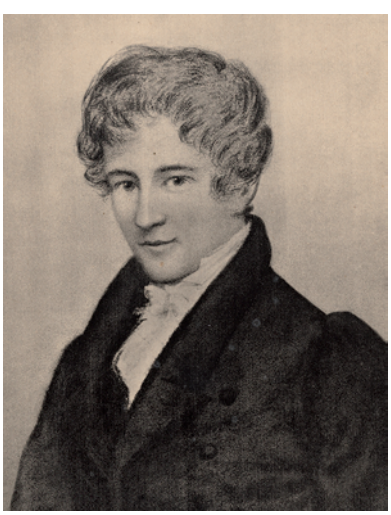

Figure 1.2 Niels Henrik Abel (1802–1829)

A good reference for all this material is Tignol (2001).

Quintic equations – Galois's work

This material relates to Book B, especially Chapters 5 and 7.

Figure 1.3 Augustin-Louis Cauchy (1789–1857)

A crucial breakthrough that Abel exploited was Cauchy's study of permutations. Augustin-Louis Cauchy (Figure 1.3) was the dominant French mathematician of his day, a gifted if egocentric thinker with the ability to work on any and every branch of mathematics and switch quickly between them, who is probably best remembered for his work in creating rigorous real and complex analysis.

In a long paper of 1815, Cauchy had shown that the number of values a rational expression in n variables can take must be greater than or equal to the largest odd prime that divides n unless it is equal to 2. In only slightly more modern language, Cauchy showed that the size of an orbit of a subgroup of the group S_n is either greater than or equal to p, the largest odd prime dividing n, or it is 2. The size of the orbit is the index of the stabiliser of a point of that orbit, and so the size of the stabiliser is either less than or equal to $\frac{n!}{p}$ or it is $\frac{n!}{2}$. This rules out subgroups of index 3 or 4 in S_5, so it has no subgroups of orders 40 or 30.

This connects neatly to what has been done before, because an expression in n variables that takes precisely k values under the $n!$ permutations of the variables is one that is fixed by a subgroup of order $\frac{n!}{k}$ of the full permutation group. So the non-existence of a group of order 30 in the permutation group on five symbols corresponds to the non-existence of an expression in five variables that takes on $\frac{120}{30} = 4$ values. Similarly, there is no expression taking only $\frac{120}{40} = 3$ values. This in turn means that the roots one can extract when solving a quintic equation are either square roots or fifth roots, and this was the decisive simplification that enabled Abel to do what Lagrange had not been able to do.

The problem now became identifying which polynomial equations of degree 5 or more are solvable by radicals. There are many of these. Carl Friedrich Gauss – see Subsection 3.9 – had given a rich theory of the equations that give n roots of unity:

$$x^n - 1 = (x-1)(x^{n-1} + x^{n-2} + \cdots + x + 1) = 0,$$

of which the interesting one is

$$x^{n-1} + x^{n-2} + \cdots + x + 1 = 0.$$

He had shown, for example, that when $n = 19$ the equation is solved by solving a succession of one quadratic and two cubic equations (note that $2 \cdot 3 \cdot 3 = 18$). When $n = 17$, and so $n - 1 = 16 = 2^4$, the equation is solved by solving a sequence of quadratics, and so by straightedge and circle (unmarked ruler and compass) alone – the first new construction of that kind since the Greeks. Gauss was only 18 when he discovered this result in 1796.

The particular case of constructions that are possible by ruler and compass alone is discussed in Book E, Chapter 19, Section 2.

The problem of deciding which polynomials are solvable by radicals is the one solved by Évariste Galois (Figure 1.4), a French mathematician who died at the age of 20 as a result of injuries obtained in a duel.

Galois made several innovations in his work. First, he observed that although every polynomial equation has a group that permutes its roots,

The definitive edition of Galois's works is Neumann (2011), which gives the French texts with a facing English translation. For a short, accurate life of Galois, see the essay about him in Rothman (1989).

different equations of a given degree (say n) may have different groups. Consider, for example, the equation for complex fifth roots of unity. It is

$$x^4 + x^2 + x^2 + x + 1 = 0. \tag{1.1}$$

Let $\omega = e^{2\pi i/5}$ be one of the roots; then the other roots are ω^2, ω^3 and ω^4. In this instance, every permutation of the roots is a power of the permutation ρ that takes ω to ω^2:

$$\rho(\omega) = \omega^2.$$

To see this, observe first that every permutation of these roots is known when it is known on ω. Next, we know that

$$\rho(\omega^2) = (\omega^2)^2 = \omega^4, \quad \rho(\omega^3) = (\omega^3)^2 = \omega^6 = \omega$$
$$\text{and} \quad \rho(\omega^4) = (\omega^4)^2 = \omega^8 = \omega^3,$$

because $\omega^5 = 1$. So we can write ρ in cycle notation as $(\omega^1, \omega^2, \omega^4, \omega^3)$. Now observe that ρ^2 can be written in cycle notation as $(\omega^1, \omega^4)(\omega^2, \omega^3)$, and that ρ^3 can be written in cycle notation as $(\omega^1, \omega^3, \omega^4, \omega^2)$, which is also ρ^{-1}. So this equation has as the group of permutations of its roots the cyclic group of order 4, not the full permutation group on four objects, which has order 24. This example was known to Gauss and Abel.

Figure 1.4 Évariste Galois (1811–1832)

Galois then adapted Lagrange's ideas to the new, more precise setting, and to state what he did we need a definition. We say that a function is *rationally known* if its numerical value is a rational function of the coefficients of the given equation and any quantities so far adjoined. Note that the decision as to whether a given expression is rationally known may change as quantities are adjoined to the field; it depends on where we are in the process of solving the equation.

When a polynomial equation is solved and its roots adjoined to the field in which the coefficients live, we obtain a new, larger field that contains the original field, and we can consider the permutations of the roots that leave the elements of the original field fixed. These permutations form a group, and we have just seen that for equation (1.1) this group is a cyclic group of order 4. But that equation is obviously special. For a typical or 'general' equation of order n with n distinct roots, one expects that the group that permutes the roots will be the full permutation group on n elements. In this case, we can also argue that any function that is rationally known is a rational function of the coefficients of the equation and, as such, is a rational function of the symmetric functions and therefore invariant under all permutations.

One key insight that Galois had was that the way to consider what happens in the process of solving a polynomial equation is to think about the group of permutations of its roots, and to watch what happens as quantities are adjoined to the original field in which the coefficients live. Different equations may have different groups, and a complete understanding of these groups would give a complete understanding of the equations themselves. This was a lesson that it took the mathematical community more than a generation to learn.

Galois now proved the following theorem.

Theorem 1.1 *Galois's First Theorem*

Given an equation (to be called the given or original equation) with roots $x_1, \ldots, x_n$, there is a group G permuting the roots of this equation with the property that

- every G-invariant function is rationally known
- every rationally known function is G-invariant.

Next Galois considered what happens when an equation is solved by solving a succession of equations of lower degree. If you look back at the solution of the cubic equation, you will see that we start with $x^3 + px + q = 0$ and proceed in stages as follows. It is best to assume that we have already adjoined a cube root of unity to the field of rational numbers (this is a number ω such that $\omega^3 = 1$).

1. We calculate $\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3$, which is a rational number.
2. We calculate $\sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}$, which we do not expect to be a rational number. We'll call one of the resulting square roots σ and the other one $-\sigma$. We are going to use σ, so we now say it is rationally known, meaning that we are now working in the field of the rational numbers with σ adjoined.
3. We calculate $-\frac{q}{2} \pm \sigma$, which does not take us to a larger field.
4. We calculate $u = \sqrt[3]{-\frac{q}{2} + \sigma}$ and $v = \sqrt[3]{-\frac{q}{2} - \sigma}$, as well as $\omega u, \omega^2 u, \omega v$ and $\omega^2 v$, where ω is a cube root of unity. This takes us from the field of the rational numbers with σ adjoined to a bigger field with these cube roots adjoined.
5. We observe that elements of this bigger field, namely $u + v$, $\omega u + \omega^2 v$ and $\omega^2 u + \omega v$, solve the cubic equation.

So, said Galois, solving an equation by radicals is a matter of steadily enlarging the field of the coefficients by adjoining pure roots (square roots, cube roots and, generally, nth roots). As the field of quantities that are rationally known is enlarged by each successive adjunction, one looks to see what the group of the equation now is. One can also be more imaginative and imagine adjoining roots of other equations, which he called auxiliary equations, and looking to see what happens.

To state what happens, Galois claimed the following theorem, which unfortunately illustrates very clearly how badly he explained his ideas (Neumann, 2011, p. 119).

Theorem 1.2

If one adjoins to a given equation the root r of an irreducible auxiliary equation, one of two things will happen: either the group of the equation will not be changed, or it will be partitioned into p groups each belonging respectively to the proposed equation when one adjoins to it each of the roots of the auxiliary equation; these groups will enjoy the remarkable property that one will pass from one to another by operating on all the permutations of the first with one and the same substitution of letters.

His idea is that if you regard a new number r as known then either you are no better off, or you are. You are better off if the group of symmetries of the equation shrinks, in the same way that Lagrange thought he was getting nearer to finding a solution every time he found an expression with an unexpectedly small number of symmetries.

Galois called an arrangement of n symbols a *permutation*, and a rearrangement of them (what we would call a permutation) a *substitution*. Such, eventually, was the success of Galois's ideas that finite groups were for a long time called substitution groups. But his terminology is the least of our problems. Doubtless because of the circumstances in which Galois wrote the memoir he was working on, he struck out the statement that the number p is the degree of the auxiliary equation, and failed to state that the partition is into j groups, where j divides p. Moreover, here 'group' simply means subset: one gets a subgroup of the original group (here called H) and its cosets. All the elements (Galois wrote *permutations*) in each coset are obtained from the first one by a single multiplication – this is how we obtain gH from H.

He did, however, sketch a proof, which is more than can be said for the rest of his memoir. He went on to state the next theorem as follows.

Theorem 1.3

If all the roots of an auxiliary equation are adjoined, each of the above 'groups' has the same elements.

Here he could only note 'One will find the proof'. The next theorem was as follows.

Theorem 1.4

Adjoining the *numerical* value of a certain function of the roots reduces the group of the equation to those elements that leave this function invariant.

Galois now came to his fundamental question: when is an equation solvable by simple radicals? By this he meant, by adjoining a sequence of nth roots for various n.

Certainly, an equation solvable by simple radicals is such that the group can be diminished to a group with a single element by adjoining a suitable sequence of roots of prime degree. Conversely, Galois said, if the group breaks up in this fashion then the equation is solvable by simple radicals. The reason is that a simple radical is a root of an equation of the form $x^p - a = 0$, where a is rationally known, and the group of permutations of the roots of the equation $x^p - a = 0$ has a particular structure that in turn forces a structure on the group of all permutations of the roots of the original equation.

Galois did try to explain this by indicating how this analysis held up in the case of quartic equations. The 'group' of the equation contains 24 *substitutions* (in our language, the 24 permutations of four objects) and, on adjoining a (suitable) square root, decomposes into two 'groups', each of 12 *substitutions*. Galois listed the elements of one of these 'groups': it is identical with what we call A_4. This group splits into three 'groups', in line with Galois's theorems on adjoining a radical of the third degree (a cube root). One of these is the group that came to be called the Klein group, $V = \{e, (12)(34), (13)(24), (14)(23)\}$, and it splits into two 'groups': $\{e, (12)(34)\}$ and $\{(13)(24), (14)(23)\}$ – notice that this second 'group' is really just a coset of the first. Adjoining a square root isolates the 'group' $\{e, (12)(34)\}$, which is reduced to the identity by adjoining a square root.

Galois concluded his memoir by investigating irreducible equations of prime degree, p. Recall that the process of solving an equation by radicals reduces the group of the equation in successive steps. Galois's analysis now proceeded by working back up from the trivial group (corresponding to the complete factorisation of the equation) to the original group.

Consider the group of an irreducible equation of prime degree p that is solvable by radicals. Galois showed that either the equation is unaffected by the adjunction of a radical or it is reduced completely. Therefore, in the reduction process, the last group reached before the group with only one element must be a group of p elements, and (as Cauchy had shown) this means the group must be the cyclic group of order p.

What about the next group up (G, say)? Galois showed that it must be made up of a certain number of 'groups' having the same 'substitutions' as the cyclic group of order p (which makes sense if we allow that Galois is alluding to the corresponding coset decomposition – the same transformations (substitutions) are applied to different lists (or permutations) of the roots). From this, by an ingenious argument, Galois deduced that the next group in the decomposition must consist of elements of the form

$$x_k \mapsto x_{ak+b},\ a, b \in \{0, 1, 2, \ldots, p-1\} \text{ modulo } p.$$

Then Galois showed that the whole group of the equation must be the group G just found, and therefore that a polynomial equation of prime degree is solvable by radicals if, and only if, its group is of the form of G.

The crucial property of this group is that it acts transitively on the roots. Moreover, no substitution of the form $x_k \mapsto x_{ak+b}$ other than the identity fixes two elements, so once two roots are known rationally so are the rest, and the group of the equation reduces to the group with one element. Therefore, for the given equation to be solvable by radicals, it is necessary and sufficient that once any two roots are given, all the others are rationally known.

In the case of a quintic equation, Galois's result implies that the largest 'group' of an equation solvable by radicals is the group of order 20 consisting of the elements

$$x_k \mapsto x_{ak+b},$$

where a and b are integers modulo 5 and b is not congruent to zero. You might like to show that in this case G is generated by $(1, 2, 3, 4, 5)$ and $(2, 3, 5, 4)$.

Therefore the general quintic equation, which has group S_5, is not solvable by radicals.

Galois's letter to Chevalier

The night before Galois went out to fight his duel, fearing that he would be killed, he wrote a hasty note to his friend Auguste Chevalier. This, as you will see, is the first place where he drew attention to the importance of adjoining roots that reduce the group of the equation to a normal subgroup. He wrote (Neumann, 2011, p. 85):

> According to Propositions II and III of the first memoir one sees a great difference between adjoining to an equation one of the roots of an auxiliary equation or adjoining them all.
>
> In both cases the group of the equation is partitioned by the adjunction into groups such that one passes from one to another by one and the same substitution; but the condition that these groups should have the same substitutions does not necessarily hold except in the second case. That is called a *proper decomposition*.
>
> In other words, when a group G contains another H, the group G can be partitioned into groups each of which is obtained by operating on the permutations of H with one and the same substitution, so that
>
> $$G = H + HS + HS' + \cdots .$$
>
> And also it can be decomposed into groups all of which have the same substitutions, so that
>
> $$G = H + TH + T'H + \cdots .$$

Here Galois presented the left and right coset decompositions of a group G with respect to a subgroup H. He did not say that two cosets are either the same or distinct, but he immediately went on to make the crucial

observation that the important case is when the left and right coset decompositions agree: every right coset is a left coset and every left coset a right coset. His word for this was 'proper', and it did not last. We say that when we have such a decomposition, the subgroup H is a normal subgroup of G. In his words:

> These two kinds of decomposition do not ordinarily coincide. When they coincide the decomposition is said to be *proper*.
>
> It is easy to see that when the group of an equation is not susceptible of any proper decomposition one may transform the equation at will, and the groups of the transformed equations will always have the same number of permutations.

This is not at all easy to see, but the key point that the behaviour of the group of the equation will answer all your questions about how to solve the equation itself is clear. Galois went on (Neumann, 2011, p. 87):

> When, on the contrary, the group of an equation is susceptible of a proper decomposition, so that it is partitioned into M groups of N permutations, one will be able to solve the given equation by means of two equations: the one will have a group of M permutations, the other one of N permutations.

Here Galois's confusion of subgroup and coset gets in the way. If we look back at the treatment of the general cubic equation, we can see that we started with the field of rational numbers, $\mathbb{Q}$, to which we had adjoined a cube root of unity, ω – which gives us the field $\mathbb{Q}(\omega)$. Then we first solved a quadratic equation by taking a pure square root and adjoining σ, and then solved two pure cubic equations by taking their cube roots and adjoining what we called u and v. The resulting field contains the solutions of the cubic equation. So we have formed a succession of fields

$$\mathbb{Q}(\omega) < \mathbb{Q}(\omega)(\sigma) < \mathbb{Q}(\omega)(\sigma)(u, v).$$

What Galois said in this case was that there is a group of six elements that permutes the roots of this cubic, but it breaks up into a subgroup, H, of three elements and the coset of this subgroup in the bigger group. In this setting, Galois's numbers M and N become $M = 2$ and $N = 3$, and his two 'groups' are a group of order 3 and its coset. The group H permutes the elements of the field $\mathbb{Q}(\omega)(\sigma)(u, v)$ and fixes those in the field $\mathbb{Q}(\omega)(\sigma)$. There is also a group of order 2 that permutes the elements of the field $\mathbb{Q}(\omega)(\sigma)$ and fixes the elements of the field $\mathbb{Q}(\omega)$ – in other words, the group that switches $+\sigma$ and $-\sigma$.

Galois continued:

> Therefore once one has effected on the group of an equation all possible proper decompositions on this group, one will arrive at groups which one will be able to transform, but in which the number of permutations will always be the same.
>
> If each of these groups has a prime number of permutations the equation will be soluble by radicals; if not, not.

In the case of the cubic, these numbers are 2 and 3. In the case of the quartic, these numbers are 2, 3, 2 and 2. In each case, the equation is solvable by radicals.

Galois then observed:

> The smallest number of permutations which can have an indecomposable group, when this number is not prime, is $5 \cdot 4 \cdot 3$.

In other words, the permutation group A_5 is simple; it has no non-trivial normal subgroups. It's a remarkable thing to have noticed, and Galois gave no hint of his proof.

Cauchy's theory of permutation groups

Cauchy's work was done in two phases, the first between 1812 and 1815 and the second around 1844–45.

In his paper 'Mémoire sur le nombre, etc.' (1815), Cauchy investigated how many different values a rational expression in n variables can take when its variables are permuted in every possible way. He noted that Lagrange and other writers had shown that it was not always possible to find a polynomial that took a specified number of distinct values (not even a number of values that divided $n!$, which is obviously necessary). Ruffini had shown in 1799 that it is not possible to find a function of $n = 5$ variables that takes only four values. Cauchy now ruled out the possibilities that a function of 5 variables takes only three or only four values. We have already seen that his work in the first phase helped Ruffini and Abel to prove that the general quintic equation is not solvable by radicals. Now we look at it in a little more detail.

In a much later paper (1844) Cauchy returned to the theme, for reasons that one speculates are to do with a revival of public interest in Galois's work, and treated the subject much more systematically. He called a list of the n symbols in an order an 'arrangement' of n symbols, and a transition from one arrangement to another he called a 'substitution' or permutation. He introduced the permutation notation (x, y, z), which we call the cyclic notation for permutations, for the arrangement, considered the product of two permutations and noted that forming the product of two permutations is not necessarily commutative. He looked at powers of a permutation, defined the identity permutation (written 1 and which he called a unity), and defined the degree or order of a permutation (the lowest positive power of it that is equal to 1). He then introduced the idea of similar permutations (which we call 'conjugate', but Cauchy reserved that word for another concept). Next he introduced the idea of a group defined (we would say 'generated') by a set of permutations – which he called a system of *conjugate* substitutions – and proved that (Cauchy, 1844, p. 207):

> The order of a system of conjugate substitutions in n variables is always a divisor of the number N of arrangements that one can make with these variables.

Here $N = n!$. Cauchy also proved that the order of any element of this group divides the order of the group.

Cauchy was aware of the concept of a normal subgroup, although he did not refer to Galois's proper decompositions, and in his paper of 1844 he established a partial converse to Lagrange's Theorem, showing that if a prime p divides the order of a group then there is an element of order p in that group and so a subgroup of order p (the condition that p be prime is necessary: there is no subgroup of order 6 in A_4, although 6 divides 12, which is the order of A_4).

Wantzel's resolution of two impossible classical problems

We discuss these theorems in Book E, Chapter 19, Section 2.

We now look at the discovery of the (nowadays) celebrated impossibility theorems that follow from Galois Theory: it is impossible by straightedge and compass to trisect an arbitrary angle or to find the side of a cube that has double the volume of a given cube. These are two of the so-called three classical problems of Greek geometry. The third is the problem of squaring the circle: find by straightedge and compass alone a square that is equal in area to a given disc (i.e. construct a segment of length $\sqrt{\pi}$). It is deeper and different and cannot be discussed here – to this day, squaring the circle is a synonym for impossibility.

The three classical problems have a complicated history. On the one hand, they were not distinguished in classical times as *the* three outstanding problems. Another, for example, asked for the construction of a regular heptagon. Nor, on the other hand, is it clear what made them difficult – or, indeed, left them unsolved. It was known by the time of Euclid that the cube can be duplicated by a construction involving the intersection of a hyperbola and a parabola. A construction for trisecting an arbitrary angle was found by Archimedes that involves sliding a marked straightedge around while ensuring that it passes through a given point. We have no source that specifies a restriction to the use of a straightedge and compass alone; the implication seems to have been drawn in modern times from the absence of any other kind of 'instrument' in Euclid's *Elements*. Nor is it clear whether the problems ask for a proof that there is an angle one-third the size of a given one (or that there is a cube twice the size of a given one) or only for a precise constructive procedure.

It is not difficult to persuade yourself that a straightedge and compass construction can only involve rational operations on known lengths and the extraction of square roots. Duplicating the cube and trisecting the angle involve cubic equations and cube roots, and the consensus by 1800 was that straightedge and compass constructions will be inadequate to the task. That answer is correct, and the first mathematician to establish this result was a French mathematician, Pierre Laurent Wantzel, who published his demonstration in 1837, when he was 23, in a mainstream journal, the *Journal des mathématiques*. Wantzel promptly vanished into obscurity, however, from which he has only just been rescued by the Danish historian of mathematics Jesper Lützen (see Lützen, 2009). Prior to Lützen's work there were very few references. There was one earlier mention, by Julius Petersen (a Danish mathematician) in 1877, but no one picked it up except Felix Klein – an influential mathematician of the late nineteenth century after whom the Klein group is named – and he dropped the reference to

See Wantzel (1837).

Wantzel by name. The first historian of mathematics to mention Wantzel by name was the North American historian Florian Cajori in 1918.

Wantzel's paper was in four parts.

1. He formulated the geometric straightedge and compass procedure as an algebraic one.
2. He deduced that the solution of a geometric problem under this restriction appears as a root of an equation of degree 2^n.
3. He claimed that this equation is irreducible (his argument was inadequate, but Pedersen and Klein repaired it easily enough).
4. However, he went on, duplicating a cube and trisecting an angle lead in general to irreducible cubic equations, which therefore cannot be solved by a sequence of quadratics.

Why did Wantzel not become famous? Lützen argues that it may be that people thought the result was already known, because Gauss had claimed in the *Disquisitiones Arithmeticae* (1801) that the only polygons that are constructible by straightedge and compass have $2^n p_1 p_2 \cdots$ sides, where the p_i are distinct Fermat primes. Therefore the regular 9-gon is not constructible; therefore the angle of 40° that appears at the centre of the 9-gon is not constructible; this is one-third of the (constructible) angle of 120°; therefore angles cannot, in general, be trisected by ruler and compass. Admittedly, for once Gauss had not supported his claim with a proof, but no one seems to have doubted that Gauss knew how to prove it – so for anyone persuaded of Gauss's claim, Wantzel was saying nothing new except for providing a proof.

Lützen also argues that the result may not have seemed interesting. People of this opinion could argue that it merely confirmed what had been suspected all along: two of the classical problems are, after all, insoluble. The result leads nowhere, and on the other hand, the equations are indisputably solvable – why care about a requirement that they be solved in a particular and frankly old-fashioned way? It would seem that the taste for impossibility theorems came later.

Lützen could find only one article about Wantzel written by someone who knew him, from which we learn that Wantzel did not fulfil his early promise. Ground down by a heavy teaching load, which he attempted to soften with coffee and opium, he died at the age of 33, his work already forgotten.

1.2 From Galois Theory to finite group theory

Galois's theory was considered very difficult in its day, and his memoir on which equations are solvable by radicals was not even published by the time he died in 1832. Then, in 1843, the mathematician Joseph Liouville announced to the Paris Academy of Sciences that he had discovered among Galois's unpublished papers a 'solution which is both correct and profound' to the problem of determining when an irreducible equation of prime degree is solvable by radicals. He also promised to publish the

memoir together with a commentary elucidating it, and explained that Galois had entrusted his manuscripts to his friend Auguste Chevalier, who had prepared an annotated version of the memoir that he had recently brought to Liouville's attention. By this time Liouville was the editor of the *Journal des mathématiques*, and the hope was that he would see that Galois's ideas finally made it into print. Liouville did indeed prepare them for publication in his journal in December 1843, only to withdraw the paper at the last minute. His reasons remain unclear, but when in 1846 he published a longer account, including a fragment of another paper, the last letter to Chevalier and all of Galois's previously published papers, he notably failed to provide a commentary.

The lack of a commentary is even more suggestive because in the winter of 1843–44 Liouville had given a short course of private lectures in Paris on Galois's ideas, which the mathematician Joseph Serret attended. Unpublished notes by Liouville from this period show that he had filled in the largest gaps in Galois's sketchy exposition, but it seems he never pulled them together into a systematic exposition. Nor was Serret to do much better. He avoided the topic in the first edition of his own *Cours d'algèbre supèrieure* (Course of higher algebra) of 1849, perhaps politely leaving room for Liouville to deliver on his promise. In the second edition (1854), all he did with Galois's work was to summarise it in a few lines (on p. 4). He asked the question 'When is an equation solvable by radicals?', and replied:

> This difficult question was completely answered, at least for irreducible equations of prime degree, by Evariste Gallois [*sic*], former student of the École Normale and one of the deepest geometers [mathematicians] that France has produced. In a paper presented to the Academy of Sciences in 1831 and published in 1846 thanks to the efforts of Liouville, Gallois in effect proved the following beautiful theorem: For an irreducible equation of prime degree to be solvable by radicals it is necessary and sufficient that, given any two of the roots, the others are rationally deducible.

What Serret did contribute was a deepening appreciation of permutation group theory, which Cauchy had pioneered in 1815 and more substantially in 1845. And when, in the third edition (1866), Serret worked through Galois's memoir, he stuck closely to what Galois had done and how he had done it, noting that 'I have followed the order of the propositions adopted by Galois, but, very frequently, I had to close the gaps in the proofs' (*Cours*, vol. 2, p. 607). He did not present the reworking of Galois's propositions that was needed to give them a new life some 30 years after their author's death.

The man who did that was a French mathematician of the next generation, Camille Jordan (Figure 1.5). He wrote a 660-page book, the *Traité des substitutions et des équations algébriques* (1870), on the theory of permutation groups and polynomial theory of equations, which is a landmark in the creation of group theory.

Jordan's *Traité* is an attempt to establish a new field of algebra, and Galois Theory as Jordan presented it is only one part of that endeavour – he also applied group theory to a wide range of complicated geometrical situations in the *Traité*. Jordan also moved the approach away from solving equations and towards their associated groups. That said, his account provides the right template for putting Galois's fragmentary account together.

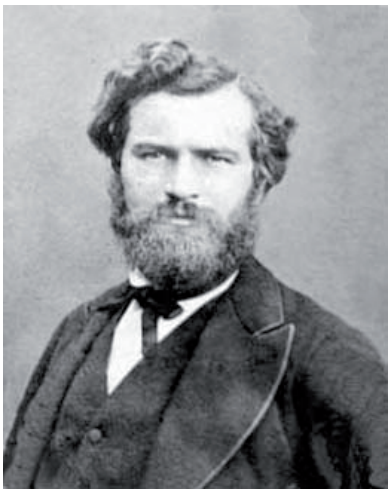

Figure 1.5 Camille Jordan (1838–1922)

Jordan followed Galois in associating to each polynomial equation a group, which he called the *Galois group* of the equation. He then showed, much as Galois had indicated, how the step-by-step process of solving an equation by successive adjunctions of roots of other polynomial equations can be understood as a series of operations on the Galois group. Moreover, Jordan had no method of finding the Galois group of an arbitrary equation except in the simplest cases, and this had the effect of making the journey from equations to groups something of a change of subject. It implied that mathematicians should loosen their grip on equations and take hold of groups instead.

Jordan's systematic theory of permutation groups was indeed highly abstract. Among the concepts he stressed, some of which he introduced, were the properties of commutativity, conjugacy, centralisers, transitivity, 'normal' subgroups (and, if only obliquely, quotient groups), group homomorphisms and isomorphisms. He was one of the first to use the Sylow Theorems in the treatment of finite groups.

The *Traité* is concerned only with finite groups, which Jordan often presented as matrix groups with entries modulo a prime number, and so it opens with some material on modular arithmetic. It then sets out the basic elementary propositions in the theory of groups: Lagrange's Theorem and Cauchy's Theorem. Lagrange's Theorem is stated for the first time in the form it retains to this day: the order of a subgroup divides the order of the group, and the proof is the coset decomposition argument. Cauchy's Theorem (**S**40 of his 1844 paper) is the partial converse: if a prime p divides the order of a group, there is an element of order p in the group.

Modular arithmetic was introduced in Book A, Chapter 3, Section 1.

Lagrange's Theorem for groups is proved in Book B, Chapter 5, Subsection 2.2.

Cauchy's Theorem is proved in Book B, Chapter 6, Subsection 3.3.

Jordan was very interested in transitive group actions: a group is n-fold transitive if, given two ordered sets of n distinct elements, there is an element of the group mapping the first set onto the other. Jordan proved the Orbit–Stabiliser Theorem: given a set of n elements of a group G, if these elements can be sent to m different places in G, and n is the order of the subgroup that leaves these elements fixed, then the order of the group is mn.

The Orbit–Stabiliser Theorem is proved in Book B, Chapter 6, Subsection 3.1.

Transitive group actions are important in the theory of polynomial equations because, as Jordan proved, the Galois group of a polynomial equation acts transitively on the roots if, and only if, the equation is irreducible. Accordingly, when one solves a polynomial equation by successive adjunction of numbers, the worthwhile adjunctions are the ones that reduce the size of the Galois group; every time a Galois group is reduced to one that acts intransitively on the roots, the polynomial equation factorises. Eventually the Galois group is reduced to the identity, the equation is completely factorised, and all the roots are therefore known.

So, in the study of irreducible quartic and quintic equations, only the transitive subgroups of S_4 and S_5 need to be considered. The Orbit–Stabiliser Theorem tells us that if a group acts transitively on a set of n elements then it has an orbit of order n, and therefore n divides the order of the group. In the case of subgroups of S_5 this tells us that a transitive subgroup contains a 5-cycle, whereas in the case of S_4 it tells us that the transitive subgroups have order $4m$ – but this allows for the possibility that we might be looking at the Klein group, which does act transitively but does not have an element of order 4.

The transitive subgroups of S_4 are $S_4, A_4, S(square)$ and V (the Klein group).

The transitive subgroups of S_5 are $S_5, A_5, G(20), S(pentagon)$ and C_5.

Next, Jordan looked in more detail at how a group breaks up into subgroups, a topic that Galois had hinted at very obscurely. Jordan was interested in a subgroup H of a group G that commuted with all the elements of G – a relationship he called 'permutable', meaning $gH = Hg$ for all g in G. (In terms of the elements involved, this says that for every $h \in H$ and every $g \in G$ there is an $h' \in H$ such that $gh = h'g$.) Jordan called G composite if it had a non-trivial subgroup of this kind and simple otherwise, and he observed that every group admits a chain of subgroups $G, G_1, G_2, \ldots$ in which G_1 is permutable with G, G_2 is permutable with G_1, and so on. He did not introduce a special word for a subgroup that 'permutes' with a larger group; the word 'normal' that is used today was introduced later by German mathematicians (we use it here in quotes).

Jordan analysed a chain of 'normal' subgroups, each one 'normal' in the one before, say

$$G = G_0 \rhd G_1 \rhd \cdots \rhd G_j \rhd G_{j+1} \rhd \cdots \rhd G_n = \{e\},$$

in terms of the orders of the quotients $\frac{|G_j|}{|G_{j+1}|}$, which he called the composition factors. In Book II, Section I, Subsection IV of his *Traité* he proved that when the chain is *maximal* – which means that it cannot be refined – the composition factors are unique up to order. It was only in 1873 that Jordan developed the theory in terms of the corresponding quotient groups, $\frac{G_j}{G_{j+1}}$, and regarded the composition factors as capturing the size of the quotient groups themselves.

Some examples may help. The composition series of S_4 is obtained from the maximal chain

$$S_4 \rhd A_4 \rhd V \rhd \mathbb{Z}_2 \rhd \{e\},$$

where V is the Klein group, and is therefore $2, 3, 2, 2$.

The composition series of S_5 is obtained from the maximal chain

$$S_5 \rhd A_5 \rhd \{e\},$$

because the group A_5 is simple, and is therefore $2, 60$.

To follow Jordan's argument that the quintic is not solvable by radicals, we need to introduce two more of his terms. He proposed to call an equation

abelian if its corresponding group is commutative – this is the origin of the term 'abelian' to refer to commutative groups. And he introduced the term *soluble groups* for the groups that characterise the equations that are solvable by radicals. His argument ran as follows.

See paragraphs 519–23 of Jordan (1870).

Theorem II For an equation to be solvable by radicals, it is necessary and sufficient that its resolution reduces to that of a sequence of abelian equations of prime degree.

Theorem III For an equation to be solvable by radicals, it is necessary and sufficient that its composition factors all be prime.

Corollary I Every group Γ contained in a soluble group G is itself solvable. For its composition factors, dividing those of G which are prime, are themselves prime.

Corollary II The general equation of degree n is not solvable by radicals if $n > 4$. For its composition factors, 2 and $\frac{n!}{2}$, are not all prime. (Jordan gave a 'proof' that the alternating group on n objects is simple when $n > 4$, but his argument is flawed, and seems to have been quietly corrected or replaced by subsequent authors.)

Corollary III The general equation of degree 3 is solvable by radicals. For its composition factors, 2 and 3, are prime.

Even before Jordan's book was published Galois's ideas had reached Germany, where they were taken up by Leopold Kronecker and Richard Dedekind, but it seems right to concentrate on Jordan's account, which was the most systematic.

1.3 Finite fields (Galois, Jordan, Steinitz)

The concept of a field has a curious history, as one might expect, because as we grow up we become very familiar with situations where we can add, subtract, multiply and divide (except by zero). The field of rational numbers does not seem to need a name or to require a new idea to become clear; it feels more like something that has always been there. Nor does much change when we learn about the fields of the real and then the complex numbers.

What slowly began to change mathematicians' opinions was the arrival of more unusual objects: the fields of algebraic numbers that arose as Galois Theory became better understood. These were explored for two reasons: the algebraic integers within a field of algebraic numbers had interesting number-theoretic properties, and rational functions in one or more variables were important in algebra and geometry. It seems that the first person to think there was now an abstract setting worthy of being given a name was Dedekind, who called a 'field' of algebraic numbers a 'Körper' in German. The word means a 'body', as does its translation 'corps' in French, and suggests an organic unity, which the English word 'field' does not – it is unclear how we came to settle on a blander term, but it was introduced by the North American mathematician Eliakim Hastings Moore in 1893.

It had also been noticed that there were some finite settings where the four basic operations of arithmetic could be carried out: the collection of integers modulo a prime. Gauss, in his *Disquisitiones Arithmeticae*, proved that the non-zero elements in such a collection are all powers of a single element, the so-called primitive root. But it was Galois, in one of the few papers he published in his lifetime, who saw more deeply.

Primitive roots are discussed in Book A, Chapter 4.

Galois claimed that if one took a polynomial f of degree ν that was irreducible over F_p, the field of p elements, and formed the collection of expressions of the form

$$a_0 + a_1 i + a_2 i^2 + \cdots a_{\nu-1} i^{\nu-1},$$

where the a_js are integers modulo p and i is a root of f, then these expressions behave like the natural numbers – by which he meant that one can add, subtract, multiply and divide with them (division by zero only being excluded). Moreover, every expression of this kind satisfies an equation of the form

$$x^{p^\nu} = x,$$

and there is at least one x of this form that satisfies no equation of the form $x^{p^n} = x, n < \nu$. This is the theorem of the primitive element: the multiplicative group of a finite field is cyclic.

See Neumann (2011, p. 65).

Later generations took this to mean that a finite field of p^ν elements can be constructed in this way that extends the field with p elements. The converse (every finite field has p^n elements and is generated by a single element of order p^n for some prime p and some integer n) was taken to be an open conjecture until it was established by Moore in 1893. It follows that up to isomorphism there is only one field of each possible order. The passage of over 60 years from Galois's death to Moore's paper is not an indication of the difficulty of the result, so much as a measure of what had to change in mathematics before fields – and finite fields in particular – came to stand out as worthy of attention.

Even by the end of the nineteenth century, fields (like rings) were still regarded as means to an end. A mathematician might have a problem in number theory or in polynomial algebra or geometry that seemed to require sophisticated handling, and might think it worth standing back from the details and trying a general argument about fields and field extensions of the following form: the basic collection of numbers or coefficients in the problem is the rational numbers; to this we adjoin some roots of polynomial equations, add some variables that themselves satisfy some equations, and so on. Textbook writers at the highest level (for example Heinrich Weber, a friend of Dedekind's) would then say there are settings in which we can add, subtract, multiply and divide (except by zero) in the ordinary way, and these are called 'fields' (other names, now abandoned, were also used). But apart from Galois's fields, they all have the rational numbers as their basic or ground field.

What catalysed a change in perception was a paper by the German mathematician Ernst Steinitz. His 'Algebraische Theorie der Körper' (The algebraic theory of fields) first appeared as an article in the major German

mathematics journal, Crelle's *Journal für Mathematik*, in 1910, and was eventually judged so important that it was reprinted as a book in 1930 with additional material by Reinhold Baer and Helmut Hasse, who hailed the paper as the starting point for many far-reaching researches in the domain of algebra and arithmetic. It marks the moment when fields themselves became objects of interest and were not merely means to other ends.

The paper opens with an introduction in which Steinitz says that his definition of a field is the same as Weber's: a system in which elements can be added, subtracted, multiplied and divided (only division by zero is excluded), in which the associative and commutative laws for addition and multiplication are obeyed, and in which multiplication is distributive over addition. He went beyond Weber in claiming that it was not Galois Theory, but the field concept itself that was now to be taken as central, so the aim of the paper was to survey all possible fields.

Fields are discussed in Book C, Chapter 11 and Book E, Chapters 18 and 19.

Steinitz began with the basic concepts: isomorphisms of fields, subfields, extension fields, adjunction of elements and integral domains. He then looked for the simplest possible fields, which he called *prime fields* and defined as fields that have no subfields. He proved that every field has a unique prime subfield, and that either there is a prime number $p > 0$ such that $px = 0$ for every x in the subfield, in which case the *characteristic* of the field is p, or the only number with this property is 0, in which case the characteristic of the field is 0.

Then Steinitz looked for extensions of the prime field. There are simple extensions, when a single element is adjoined, and these may be algebraic or transcendental. Algebraic extensions are finite-dimensional, and form the subject matter of Galois Theory. The rational functions in a variable x over a field K form a field $K(x)$, and if a polynomial $f(x)$ of degree n is irreducible over the field then $\dfrac{K(x)}{f(x)}$ is also a field; it is an algebraic extension of the ground field K of degree n. The new field is said to be obtained from the old one by adjoining all the roots of the polynomial $f(x)$.

Steinitz was at his most original and insightful when it came to dealing with the differing properties of fields of zero characteristic, and those of characteristic p. He gave a thorough account of fields of finite characteristic and rewrote the theory to include them systematically, with a discussion of the novel issues that arise when working with a field of characteristic p.

Amusingly, given the thrust of Steinitz's original presentation, when Baer and Hasse reissued the paper as a book they provided additional material on Galois Theory. They presented it in terms of a ground field K and a normal algebraic extension field L (a normal extension was defined by Steinitz as one with the property that an irreducible polynomial over K either remains irreducible over L or factors completely). In their hands, Galois Theory became the study of how fields intermediate between K and L relate to subgroups of the Galois group G of the extension, which is defined as the automorphisms of L that leave K pointwise fixed. Polynomials, and solvability by radicals, finally disappeared from the theory.

2 The origins of topology and metric spaces

2.1 The importance of Fourier series

Topology, and especially the study of metric spaces, grew out of analysis, and within analysis the dominant problem was the behaviour of sequences (of numbers or, more usefully, of functions) when going to a limit. Typically, one would have a sequence of functions that, formally at least, seemed to have a limit – for example, a sequence of polynomials $\sum_{j=0}^{n} a_j x^j$ with the power series $\sum_{j=0}^{\infty} a_j x^j$ as its limit as $n \to \infty$. Did the limit of this sequence of functions make sense – did it define a function? If so, what could be said about it?

Although it was known in the eighteenth century that the method of successively approximating a solution was often productive, the tools for bringing questions such as the one above into sufficient focus to be answerable were only developed in the early nineteenth century. By then there were many factors pushing mathematicians towards the epsilon-delta type of analysis that was finally introduced by Cauchy in the 1820s. Whenever functions were handled by their Taylor series approximations, or differential equations were solved by power series methods, the need arose to be sure that cutting off the series at a certain point was reliable. It was well known that some series make numerical sense only for a restricted range of the variable: for example

$$1 - x + x^2 - x^3 + \cdots ,$$

which equals $\frac{1}{1+x}$ if, and only if, $|x| < 1$. And, indeed, it was widely agreed that while the calculus of Newton and Leibniz was somehow right, there was no truly satisfactory explanation of why this was so.

Cauchy's work gave the basis of a new approach. He did not appeal to dynamical arguments in which points move along curves and reach limiting positions, and he did not appeal to geometrical arguments about tangents. Instead, he worked with an arithmetical approach based on inequalities, and did so much more systematically than anyone had done before. This had the virtue of making the calculus applicable to problems in mechanics and geometry without the risk of a vicious circle, and it steadily provided a set of procedures for proving what had hitherto been accepted largely on intuitive grounds. With Cauchy's analysis it became possible to prove the Intermediate Value Theorem, to state and prove the Mean Value Theorem, to prove that a function continuous on an interval is integrable there, to prove the fundamental theorem of the calculus, to establish numerous convergence tests for sequences and series; and so to resolve most of the problems that had begun to crop up whenever the calculus was used.

Fourier series

The next challenge came from one of the best ideas ever introduced into applied mathematics: Fourier series. Joseph Fourier argued for many years that the problem of the diffusion of heat through a solid could be solved by regarding the temperature as a function of time and position, expressing the flow of heat by means of a specific partial differential equation that he wrote down, and writing the solutions in the form of series involving the trigonometric functions. He eventually succeeding in publishing his ideas in his *Théorie analytique de la chaleur* (*The Analytical Theory of Heat*) in 1822.

In the simplest case, the solid body might be a slab heated at its base, so Fourier was interested in the temperature at a point as a function of time and the height above the base. He argued that the partial differential equation implies that the temperature is given by a possibly infinite sum of terms of the form $f(x)g(t)$, when $f(x) = \cos(mx)$ and $g(t) = e^{-mt}$, and so he looked at series of the form

$$ae^{-y}\cos x + be^{-3y}\cos 3x + ce^{-5y}\cos 5x + de^{-7y}\cos 7x + \text{etc.} \qquad (2.1)$$

Other Fourier series may involve the sine function exclusively, or both sines and cosines. This one involves only cosines because of other assumptions Fourier made about the shape of the slab.

subject to the boundary condition at the base that

$$1 = a\cos x + b\cos 3x + c\cos 5x + d\cos 7x + \text{etc.} \qquad (2.2)$$

The coefficients $a, b, c, d, \ldots$ have to be determined, and the series shown to converge. Fourier claimed that the coefficients were given in this case as

$$a = \tfrac{4}{\pi}, \quad b = \tfrac{-4}{\pi}\cdot\tfrac{1}{3}, \quad c = \tfrac{4}{\pi}\cdot\tfrac{1}{5}, \quad d = \tfrac{-4}{\pi}\cdot\tfrac{1}{7}$$

by the simple formula, but took it for granted that any series of this type and obtained in this way converged. He was dealing with practical problems where formal mathematics can be supported by empirical observations, but his work raised questions for others. Did the series of his type always converge? Could any function be written as a Fourier series? It is simply a statement of fact that questions raised in the study of Fourier series did more to drive mathematicians to take up rigorous analysis than any other development in the nineteenth century.

Abel was among the first to notice that something was amiss. Fourier had given the example of the series

$$\sum_{k}^{\infty} \frac{(-1)^{k+1}\sin(kx)}{k},$$

which we will call $Ab(x)$. The graph of this function is copies of the function $y = \frac{x}{2}$ between $-\pi$ and π and discontinuous at all odd multiples of π.

A Maple graph of the first 100 terms of this series shows how good the match is between the function and the Fourier series away from the points of discontinuity (Figure 2.1), but notice how unhappy Maple gets at the points where the function will be discontinuous (the vertical lines are a Maple artefact and should be ignored).

Maple is a very commonly used computer algebra package.

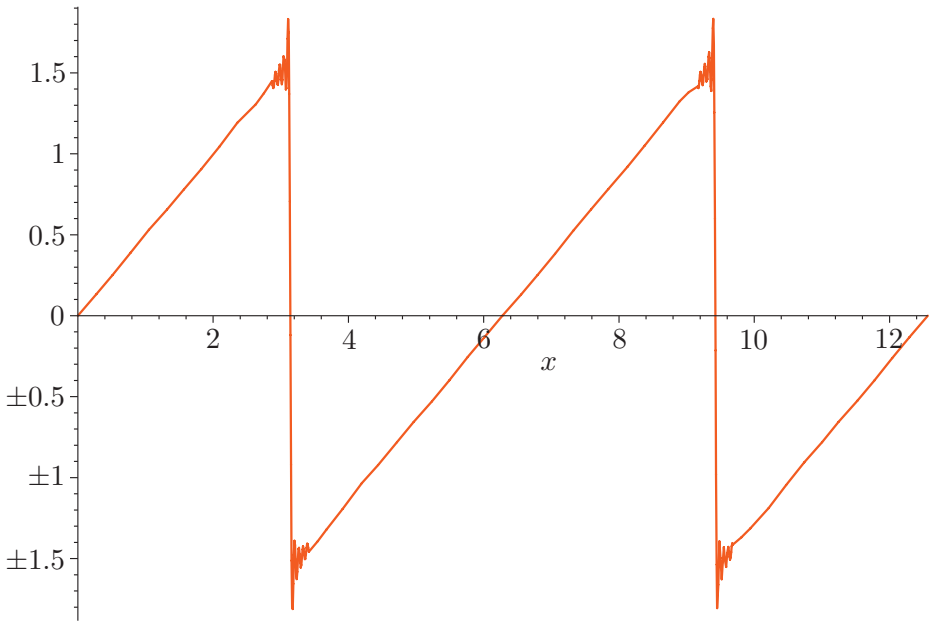

Figure 2.1 The first 100 terms of the series for $Ab(x)$

Since every term of the Fourier series is continuous (indeed, infinitely differentiable) and the series itself is visibly discontinuous at some points, we also have the interesting spectacle of a series of continuous functions not summing to a continuous function (contrary to what Cauchy thought he could prove in the 1820s).

The person who did most to start the productive investigation of these questions was Peter Gustav Lejeune Dirichlet. He wrote the Fourier series associated with a function $\phi(x)$ on the interval $[-\pi, \pi]$ as

$$\tfrac{1}{2}b_0 + b_1 \cos x + b_2 \cos 2x + \cdots + b_m \cos mx + \cdots$$
$$+a_1 \sin x + a_2 \sin 2x + \cdots + a_m \sin mx + \cdots$$

where the coefficients are given by Fourier's formulas:

$$b_m = \frac{1}{\pi} \int_{-\pi}^{\pi} \phi(\alpha) \cos(m\alpha) d\alpha, \tag{2.3}$$

$$a_m = \frac{1}{\pi} \int_{-\pi}^{\pi} \phi(\alpha) \sin(m\alpha) d\alpha. \tag{2.4}$$

He then proved, in a beautiful exercise in rigorous epsilon-delta analysis, that this series converges to the given function at every point where the function is continuous, provided that the interval $[-\pi, \pi]$ can be divided into a finite collection of sub-intervals on each of which the function is monotonic (Dirichlet's Theorem). He also established what happens if the function has a jump discontinuity at a finite number of points. Let $a \in [-\pi, \pi]$ be such a point, and suppose that

$$\lim_{x \uparrow a} f(x) = A_- \quad \text{and} \quad \lim_{x \downarrow a} f(x) = A_+,$$

where the limits are taken from below and above respectively. Dirichlet now showed that the Fourier series at $x = a$ takes the value $\frac{1}{2}(A_- + A_+)$.

This took care of the problem Abel had raised, but it opened many more. If the restriction to jump discontinuities is dropped, and a function such as

$$f(x) = \begin{cases} \sin\left(\frac{1}{x}\right), & \text{if } x \neq 0, \\ 0, & \text{if } x = 0, \end{cases}$$

which is discontinuous at $x = 0$, is admitted, then Dirichlet's Theorem has nothing to say. Functions with this kind of discontinuity are said to oscillate infinitely in the neighbourhood of the point where they are discontinuous, but there are also continuous functions that oscillate infinitely in an interval (for example, $x \sin\left(\frac{1}{x}\right)$) and they too are not covered by Dirichlet's Theorem. These functions have infinitely many maxima and minima. What can be said about them? Dirichlet himself pointed to an even worse example: Fourier's formulas for the coefficients only work if the function ϕ is integrable. What, he asked, can be said when ϕ is not integrable, and is, for example, the function that takes one value, c, when x is rational and another value, d, when x is irrational?

Uniform convergence

It is one thing to point towards work that has to be done, but quite another to find a way of doing that work. The next big advance in the study of Fourier series was made by Dirichlet's greatest student, the shy but brilliant Bernhard Riemann.

It is worth pausing here to note that at various stages in the mid-nineteenth century a great 'safety barrier' was created, which defined a class of functions with most of the intuitive properties one would expect. This barrier was formed from the concepts of uniform continuity and uniform convergence. We will define these terms for completeness, but the only point to note is that a 'fence' was erected within which functions can be proved to have a wide class of intuitively accessible properties. Attention then shifts to functions in the wild country beyond the fence.

A sequence of functions (f_n) defined on a set of points T converges *uniformly* to a function f if, for every $\varepsilon > 0$, there is an N such that $n > N$ implies $|f_n(x) - f(x)| < \varepsilon$ for every x in T. It soon became clear that if each of the functions f_n is continuous at a point t_0 in T and the f_n converge uniformly in T then the limit function f is also continuous at t_0. It was also found that the uniform convergence of a sequence of integrable functions f_n to a function f on a set T (given a suitably precise theory of the integral, which we are carefully not specifying) guaranteed not only that the limit function was integrable, but also that the sequence of integrals $\left(\int_T f_n(x)dx\right)$ converged to the integral $\int_T f(x)dx$, so term-by-term integration of a uniformly convergent series is a valid operation.

The situation for sequences of differentiable functions was a little more complicated, and we omit the details in favour of the moral of the story, which is that uniform convergence of well-behaved functions produces well-behaved limits. There is no such guarantee without uniformity. The result was that mathematicians were forced to look for uniformity if they wanted the strong intuitive results, and to work without it if they wanted to look for functions that were intelligible (because they were limits of

well-behaved functions) but not themselves well-behaved (and were perhaps continuous but not differentiable, or integrable but not continuous, or not even integrable). The second, much more arduous route was first taken by Riemann.

Non-uniform convergence and strange functions

Between 1851 and 1854 Riemann wrote his post-doctoral work under Dirichlet's direction. (This gave him the necessary qualification to teach in a German university.) He prefaced his final essay with a lengthy historical account of the history of Fourier series, going back to Euler. He then argued that although functions which do not satisfy Dirichlet's conditions are unlikely to occur in nature, they nonetheless have a 'very close connection with the principles of infinitesimal calculus, and can serve to bring greater clarity and rigour to these principles'. Moreover, he had found that such functions occur in parts of analytic number theory, so they were important in pure mathematics.

Riemann first showed how to define integrability in such a way that certain discontinuous functions could be integrable. Then he turned his attention to functions with infinitely many points of discontinuity, and gave a provocative example. He first defined the function $r(x)$ as follows. For each real number x,

$$r(x) = \begin{cases} x - m(x), & \text{if } x \neq \frac{n}{2}, \\ 0, & \text{if } x = \frac{n}{2}, \end{cases}$$

where $m(x)$ is the integer minimising the expression $|x - m(x)|$, and n is odd. So the function $r(x)$ takes values between $-\frac{1}{2}$ and $\frac{1}{2}$, and is zero where $x = \frac{n}{2}$ (where n is an odd integer). Its graph is illustrated in Figure 2.2 (again ignore the vertical lines).

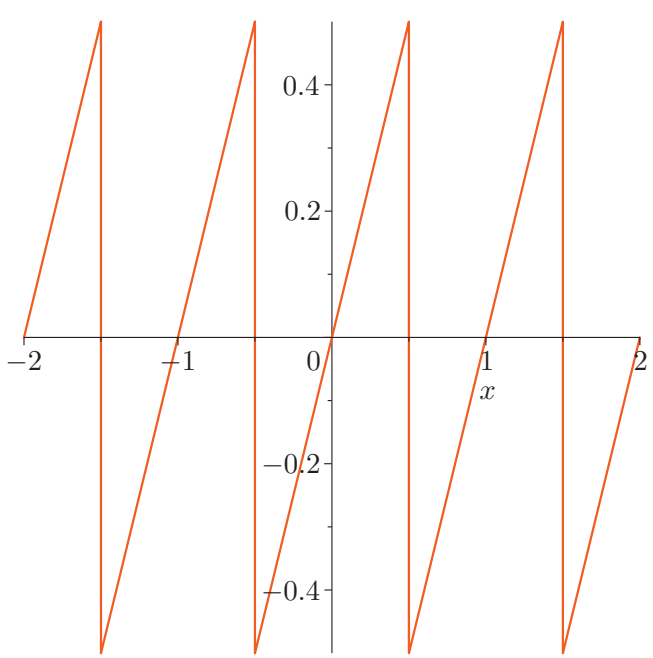

Figure 2.2 Riemann's function $r(x)$ on the interval $(-2, 2)$

Riemann then defined the function

$$f(x) = r(x) + \frac{r(2x)}{2^2} + \cdots + \frac{r(nx)}{n^2} + \cdots .$$

This function is discontinuous at every point $x = \frac{m}{2n}$, where m and $2n$ are relatively prime; this is a dense set of rational numbers. At these points, the left- and right-hand limits of the function f are

$$f(x_+) = f(x) - \frac{1}{2n^2}\left(\sum_k \frac{1}{(2k+1)^2}\right) = f(x) - \frac{\pi^2}{16n^2}$$

and

$$f(x_-) = f(x) + \frac{1}{2n^2}\left(\sum_k \frac{1}{(2k+1)^2}\right) = f(x) + \frac{\pi^2}{16n^2}.$$

The next Maple graph (Figure 2.3) shows the first 150 terms of the sum:

$$f(x) = r(x) + \frac{r(2x)}{2^2} + \cdots + \frac{r(150x)}{150^2}.$$

Again, ignore the vertical line that Maple has supplied where the function has a jump.

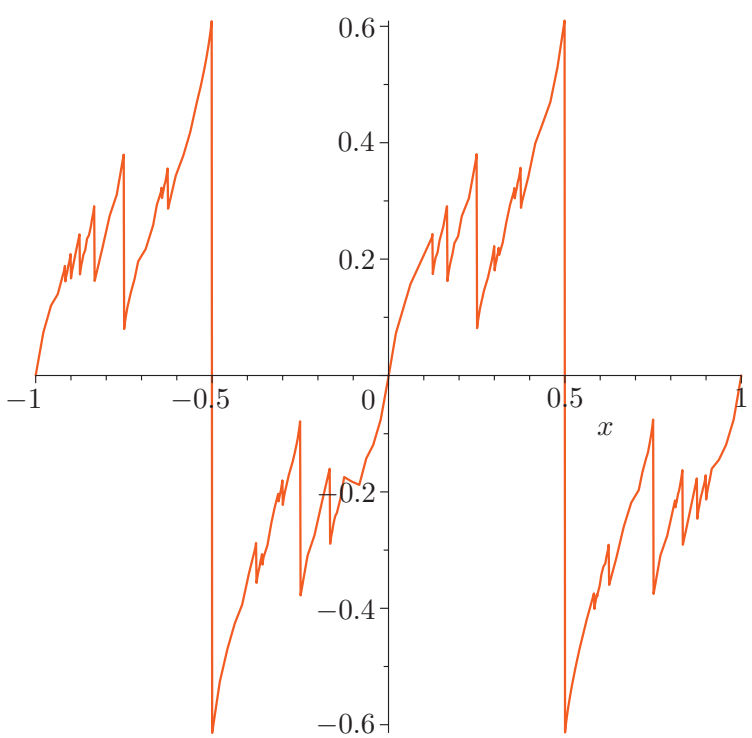

Figure 2.3 The first 150 terms of the sum defining Riemann's function

One can think of the integral $\int_a^x f(t)dt$ as x goes from a to, say, b as calculating a moving average. By the Intermediate Value Theorem, $\int_a^x f(t)dt = (x-a)g(x)$, where $g(x)$ is the average value of $f(x)$ on the interval (a, x). So a function should be integrable if it always has an average value at every point of an interval on which it is defined. Such

functions may oscillate wildly and be discontinuous, and even fail to be defined at finitely many points, so it is reasonable that integrable functions might be discontinuous at some points and oscillate to some degree. In the example just considered, each jump affects the average value by a finite amount, but the function is integrable because the contributions of the jumps are negligible except at finitely many points.

To study functions that lie beyond the domain of Dirichlet's Theorem it is necessary to have examples, and Riemann's breakthrough was his introduction of what are called trigonometric series, which he wrote this way:

$$\Omega(x) = A_0 + A_1(x) + A_2(x) + \cdots + A_n(x) + \cdots , \tag{2.5}$$

where

$$A_0 = \tfrac{a_0}{2}, A_n(x) = a_n \cos nx + b_n \sin nx.$$

These are not necessarily the Fourier series of any function. Riemann assumed that $A_n(x)$ converged to 0 as n increased – in fact, he treated this as uniform convergence throughout the paper. The convergence is for all values of x under consideration.

Very ingenious arguments allowed Riemann to derive necessary and sufficient conditions for a function to be represented at a point by a trigonometric series Ω. In this way he opened the door to all sorts of unexpected, and largely unknown, functions. A list of his results will indicate the range of behaviour of functions that he was now able to express.

1. Functions with infinitely many maxima and minima can be integrable without being representable by a Fourier series. For example,

   $$\frac{d(x^\nu \cos \frac{1}{x})}{dx}$$

   on $0 \le x \le 2\pi$, with $0 < \nu < \frac{1}{2}$.

2. Conversely, there are functions that cannot be integrated but can be represented by a trigonometric series. For example, the function

   $$\sum_{k=1}^{\infty} \frac{r(kx)}{k}$$

   exists for every rational value of x and is representable by

   $$\sum_{1}^{\infty} \frac{s(x)}{n\pi} \sin(2n\pi x),$$

   where $s(x)$: $= \sum \sigma^d - (-1)^d$ and is taken to be positive or negative according to whether n is odd or even, and σ ranges over all the divisors of n. This series is not bounded in any interval, no matter how small, and consequently is not integrable.

   Another example is

   $$\Sigma c_n \cos(n^2 x),$$

where $c_n \to 0$ but $\Sigma c_n = \infty$. As Riemann explained (1854, **S**13), 'For, if $\frac{x}{2\pi}$ is rational with denominator m in lowest terms, then clearly the series either converges or diverges to infinity according as

$$\sum_{n=0}^{m-1} \cos(n^2 x)\,, \quad \sum_{n=0}^{m-1} \sin(n^2 x)$$

are zero or not. Both cases arise.'

3. Finally, a trigonometric series can converge for infinitely many x in any interval and yet the coefficients not tend to zero. For example

$$\sum_{1}^{\infty} \sin(n!\pi x),$$

which converges not only for every rational value of x but also for some irrational numbers, such as $\sin 1, \cos 1, \frac{2}{e}$ and its multiples, etc.

Figure 2.4 is a graph of the first four terms of Riemann's series in result (3), which converges at some points even though the coefficients do not tend to zero.

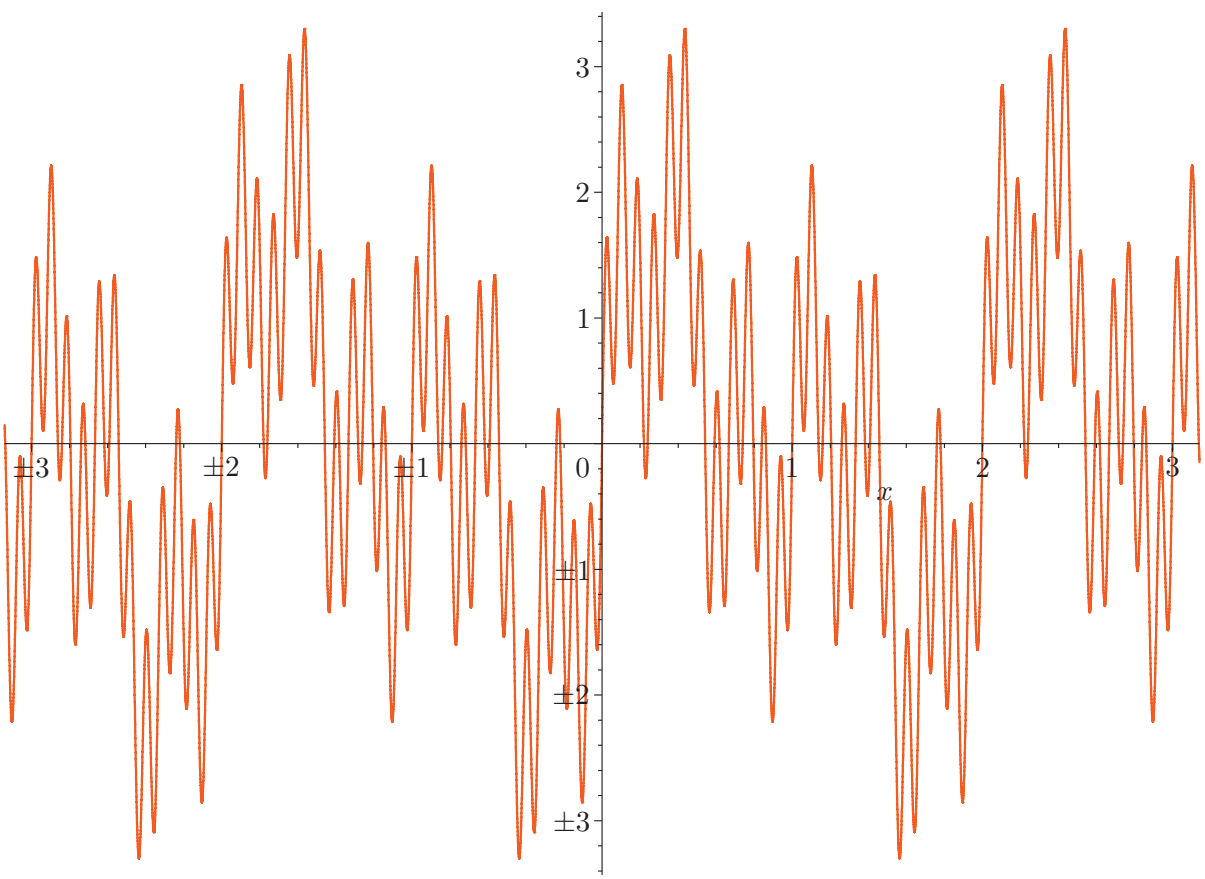

Figure 2.4 The first four terms of $\sum_{1}^{\infty} \sin(n!\pi x)$

Much of the ensuing work by a generation of analysts centred on exploring the highly counter-intuitive properties of functions that are continuous but

not differentiable. Riemann suspected, and in 1872 Weierstrass eventually proved, that there are functions continuous everywhere on an interval but differentiable nowhere. Weierstrass gave the example of the functions

$$f(x) = \Sigma_n b^n \cos(a^n \pi x) \ , \quad 0 < b < 1 \ , \ ab > 1 + \tfrac{3\pi}{2} \ , \ a \text{ an odd integer.}$$

Continuity follows from the fact that, by Weierstrass's M-test with $M_n = b^n$, the series is uniformly convergent for all x, but establishing non-differentiability everywhere is hard. The condition that the integer a be odd makes the fact that $\cos(N\pi) = \pm 1$ easier to handle. The condition that $ab > 1 + \frac{3\pi}{2}$ is what makes the tail of the series oscillate so much that the function is nowhere differentiable.

A comparison of the Maple graphs of the Weierstrass function for $a = 13, b = \frac{1}{2}$, where Weierstrass's conditions are obeyed, and for $a = 3, b = \frac{1}{2}$, where they are not, is instructive.

Figure 2.5, which shows the graph of just the first five terms of the series, powerfully suggests that differentiability is failing everywhere. By contrast, Figure 2.6, which shows the first 15 terms in the fake, suggests that differentiability is failing only at some points (albeit possibly a dense set).

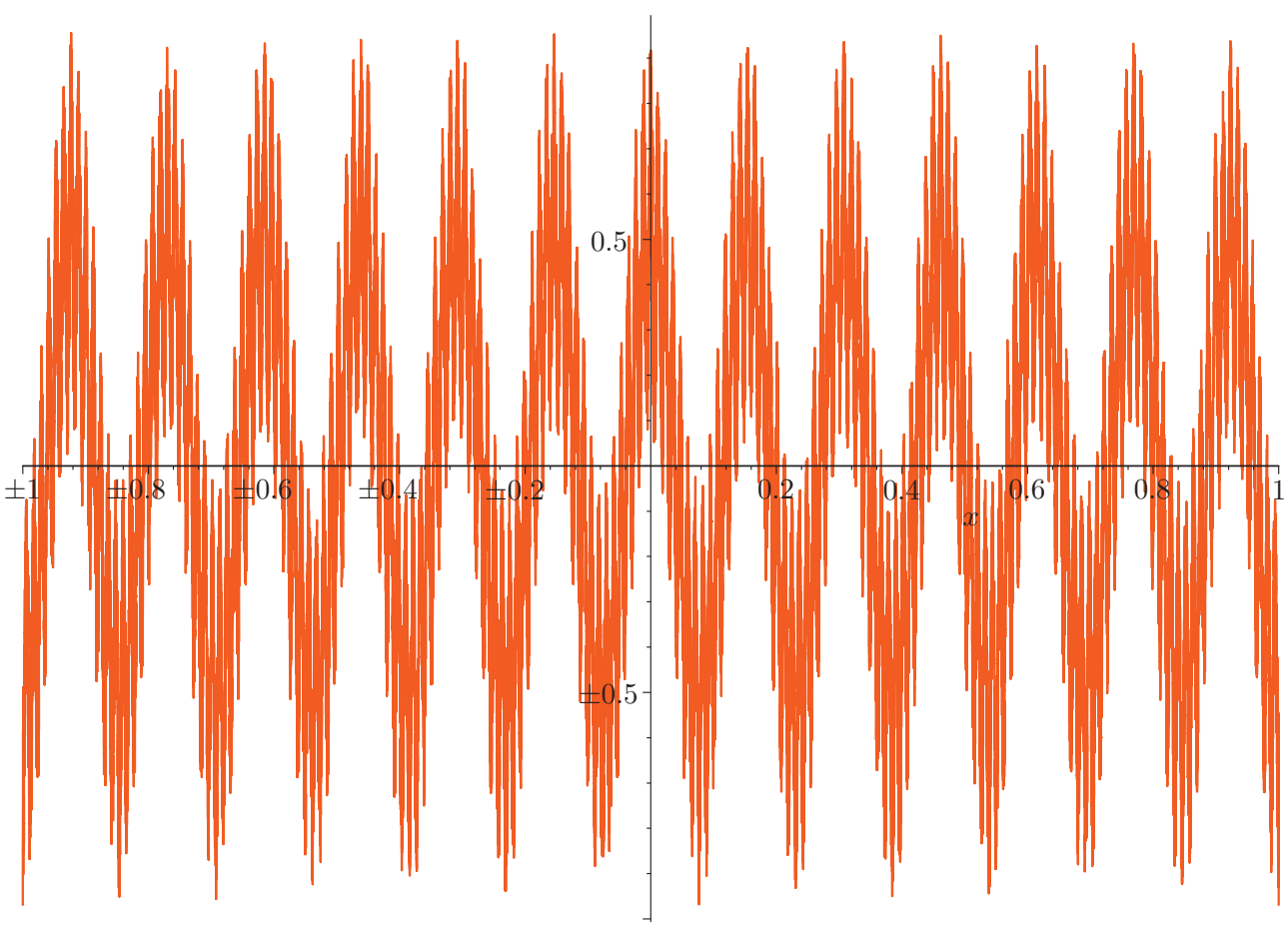

Figure 2.5 The first few terms of a Weierstrass function

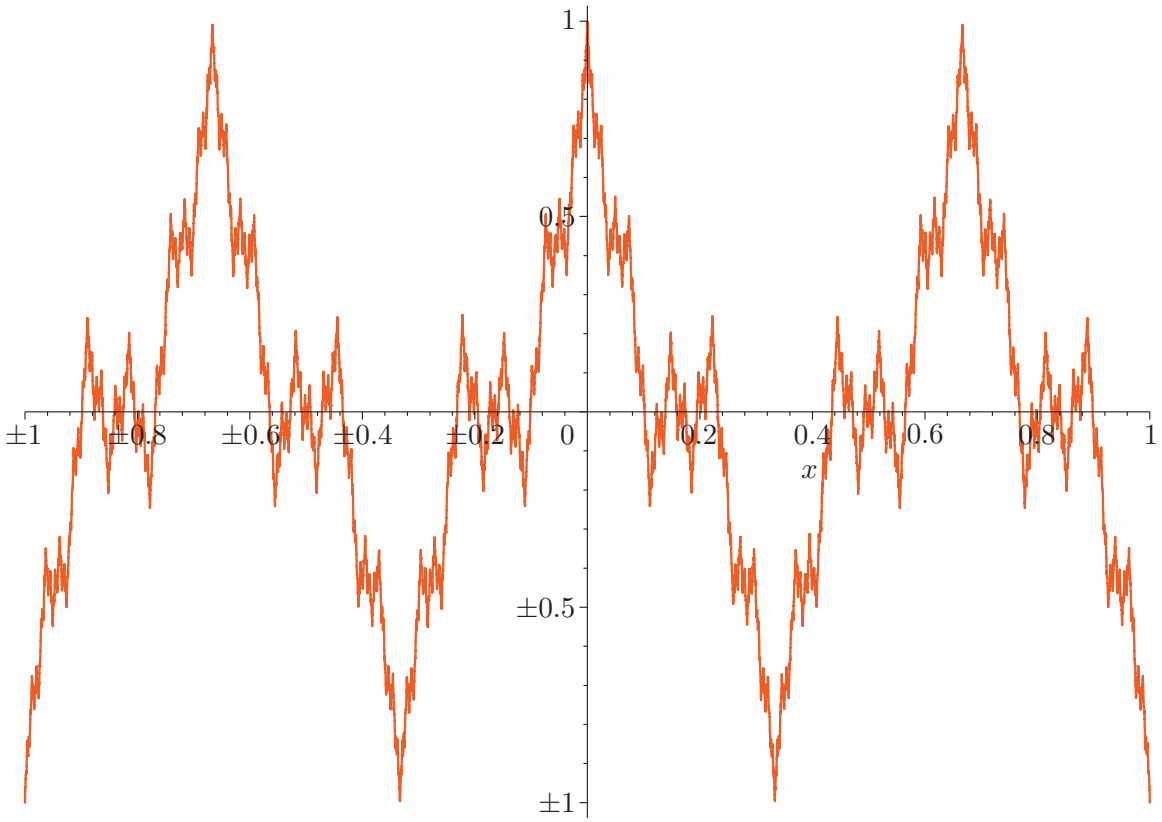

Figure 2.6 A 'fake' Weierstrass function

Cantor and point-set topology

Georg Cantor was the first to explore deeply the implications of the fact that if two integrable functions agree everywhere on an interval except at finitely many points then they have the same integral. His explorations led him to discover some of the fundamental features of point-set topology, and a decade later to his discovery of infinite sets of various sizes.

Cantor's work took off from the investigations of a colleague of his at the University of Halle in 1870. Heinrich Heine had raised the question of what can be said if the formulas for the Fourier coefficients (equations 2.3 and 2.4) are invalid, as they may be if the series is not uniformly convergent, and so term-by-term integration is illegitimate. He answered this question by giving some criteria about the set of points where continuity failed that, he thought, still allowed mathematicians to establish that the coefficients were unique. In 1872 Cantor aimed at showing uniqueness for any set of discontinuities. He investigated what can be said if a function is represented by two trigonometric series, $\sum a_j \sin jx + b_j \cos jx$ and $\sum a'_j \sin jx + b'_j \cos jx$, that converge and agree everywhere on an interval I except at a set P, where either convergence fails for at least one of them or the two trigonometric series take different values.

Cantor began with sets P that have only finitely many limit points. He called the set of limit points of a set P the derived set of P, and denoted it

P', so he began with sets whose derived set is finite. He called these sets of type 1, and proceeded inductively to define sets of higher type. A set of type 2 is one whose derived set is of type 1, and Cantor denoted the derived set of P' by $P^{(2)}$. A set of nth type is one for which its nth derived set $P^{(n)}$ is finite. At this stage Cantor had almost no examples of the sets he was talking about, and noted only that the set $\frac{1}{n} : n = 1, 2, \ldots$ is of type 1, and that the set of all the rational numbers in $[0, 1]$ is of no type. But he was able to show that the representation is unique if convergence holds except on a set of type n for some n. His proof was a modification of Heine's, extended by induction to any integer n.

By definition, a set of type 1 has a finite derived set, so there are only finitely many points p' each with the property that every neighbourhood of p' contains infinitely many points of P. These points p' need not belong to P, but the points of P somehow bunch up around them. It helps to introduce a term that Cantor was only going to introduce later: we say that a set is *nowhere dense* if its closure has an empty interior. Therefore, such a set contains no interval. It is clear that a set of type 1 is nowhere dense, and indeed that a set of type n is nowhere dense. So Cantor's theorem about the uniqueness of representation says very roughly that the representation is unique, provided that convergence holds everywhere except on a set with no interior.

Smith discovers Cantor sets

One fascinating thing to observe here is that what is at stake is what sets, in the domain of definition of a function, can be said to be negligible for the purposes of integration. However, the answer being proposed is topological. It seemed to several mathematicians in the 1870s that a set would be negligible for the purposes of integration if, and only if, it was nowhere dense. To modern eyes, this hints at a confusion, and the first to highlight it was the Oxford mathematician Henry Smith (Figure 2.7).

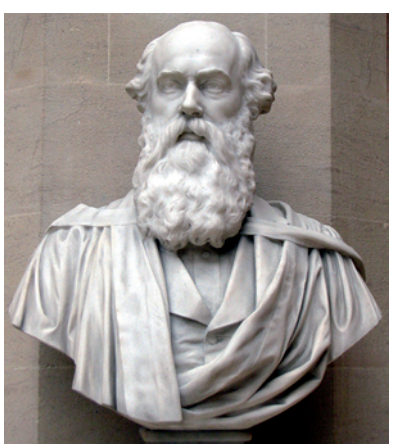

Figure 2.7 Henry John Stephen Smith (1826–1883)

Let us briefly introduce a term for a concept that was introduced in the 1870s without being given a name. If a set A is covered by a finite set $I = \{I_k\}$ of intervals, then the total length of the covering is $L_I = \sum_k |I_k|$. We are interested in the least upper bound of all finite coverings as the number of intervals in the covering increases arbitrarily because, if this quantity is zero, the set A will be negligible for the purposes of integration. Let us call such sets *zero coverable*.

Smith's first family of zero coverable sets was constructed in terms of a family of sets. The set P_1 consists of all the rational numbers of the form $\frac{1}{n_1}$, where n_1 is a integer; note that its points cluster around the point $x = 0$, which is not itself in P_1. The set P_2 consists of all the rational numbers of the form $\frac{1}{n_1} + \frac{1}{n_2}$, where n_1 and n_2 are positive integers. It is easy to check that P_1 is a subset of P_2, but there are points of P_2 that cluster around any element of P_1. The construction proceeds iteratively: the set P_k consists of all the rational numbers of the form $\frac{1}{n_1} + \frac{1}{n_2} + \cdots + \frac{1}{n_k}$, where $n_1, n_2, \ldots, n_k$ are positive integers.

You might like to confirm Smith's result that the jth derived set of $P_k, k > j$, is $P_k^{(j)} = \{0\} \cup P_1 \cup \cdots \cup P_{k-j}$. Smith showed that the sets P_k are zero coverable.

The second family of nowhere dense sets, Smith called sets 'in loose order' – see if you recognise them under another name. He defined them in these words (Smith, 1875, p. 147):

> Let m be any given [integer] greater than 2. Divide the interval from 0 to 1 into m equal parts; and exempt the last segment from any subsequent division. Divide each of the remaining $m - 1$ segments into m equal parts; and exempt the last segment of each from any subsequent division. If this operation be continued ad infinitum, we shall obtain an infinite number of points of division P upon the line from 0 to 1. These points lie in loose order ...

In more modern terms, Smith's process produces an infinite sequence of sets, each contained in the one before, and he took the set P to be all the points common to every set in the sequence. At each stage after the second, each set consists of a collection of intervals, and at each stage the new set differs from the one before because certain small intervals at the end of each interval have been removed.

The set P that Smith obtained is nowhere dense (informally, because intervals are being removed at each stage) and its complement at the nth stage is of length $(1 - \frac{1}{m})^n$, which tends to 1, so the set is zero coverable; a bounded function that is discontinuous at precisely the points of P is integrable.

Smith's set P is the first 'Cantor' set, S, after the very similar set that Cantor described in 1883. Cantor's construction resembles Smith's with $m = 3$, but at each stage the middle third of each interval is excluded. As this shows, Smith's works were not much read on Continental Europe.

Smith then changed the definitions to produce a new family of sets that are still nowhere dense but not zero coverable, so a function discontinuous on this set is not integrable. He let m be any integer greater than 2, divided the interval from 0 to 1 into m equal parts, and exempted the last segment from any subsequent division. Now he divided each of the remaining $m - 1$ segments into m^2 equal parts, and exempted the last segment of each from any subsequent division. So in this process, much smaller end intervals are being removed. The process is continued indefinitely, and the points that remain form an infinite set of points Q in the interval [0, 1]. This set Q is nowhere dense. But now at stage n the complement of the set is length $(1 - \frac{1}{m})(1 - \frac{1}{m^2}) \cdots (1 - \frac{1}{m^n})$, which does not tend to zero as n tends to infinity but instead tends to the positive limit $\Pi_n^\infty(1 - \frac{1}{m^n})$. For example, if $m = 3$ then this limit is over one-half (at about 0.560126). So the set Q is not zero coverable.

The example of a nowhere dense set in the unit interval that is not zero coverable shows that there was a confusion between a set being nowhere dense and its being zero coverable. Cantor's sets of type n are both nowhere dense and zero coverable, and this may have suggested to

mathematicians at the time that the sets which are negligible for the purposes of integration are the nowhere dense sets. What Smith showed was that this topological criterion was not the right one to use, and once this confusion was exposed the way was open for the topological concepts and the integration theory concepts to develop independently.

2.2 Potential theory and the Dirichlet principle

The topological concepts of connectedness and path-connectedness have their origins in another branch of mathematics: potential theory. The theories of gravitation and of electricity and magnetism in the nineteenth century had a common feature: it was assumed there was a force involved, which was represented by a vector. Because vectors are more complicated to handle than functions, George Green in England and Gauss in Germany argued independently in the 1830s and 1840s that the force was usually the derivative of a function. They claimed that the force $\mathbf{F}$ of, say, gravity due to a solid body could be written as $\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right)$ for a suitable function f. Green gave this function a name that it still retains: the potential function. For gravity and electricity, the potential function is harmonic:

$$\triangle f = \nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = 0.$$

Very quickly mathematicians became interested in proving the existence of such a function under very general conditions, corresponding to different distributions of mass or electric charge. Physicists, of course, were happy to assume potential functions always existed. The problem is different in two and three dimensions, and the simpler but very productive two-dimensional problem became this: given a domain D in the plane bounded by a closed curve C, and given a continuous function f on C, find a function φ defined on D that is harmonic on the interior of D and is equal to f on C. The idea is that the function φ represents some distribution of mass or electricity on C, and the function f is to be the corresponding potential function.

The problem proved very difficult to solve, despite the physicists' confidence. Green had claimed in 1833 that the three-dimensional problem can always be solved, and his argument was that the required function is the one that minimises the integral

$$\int_{vol} \left(\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + \left(\frac{\partial f}{\partial z}\right)^2 \right) dx\, dy\, dz.$$

William Thomson (later Lord Kelvin) reaffirmed the principle himself in 1847, as did Dirichlet during lectures in 1856–57 at Göttingen (which were published posthumously in 1876).

In Dirichlet's opinion, the sought-for function u could seldom be found explicitly: 'we can only speak of an existence proof for it. The latter presents no difficulty.'

He went on (Dirichlet, 1876, pp. 127–8):

> For every bounded connected domain T there are clearly infinitely many functions u continuous together with their first-order derivatives, for x, y, z which reduce to a given value on this surface. Among these functions there will be at least one which reduces the following integral $U = \int_{vol}[(\frac{\partial u}{\partial x})^2 + (\frac{\partial u}{\partial y})^2 + (\frac{\partial u}{\partial z})^2]$ extended over the domain T, to a minimum; it is evident that this integral has a minimum since it cannot become negative. We can now show the following:
>
> 1. Every such function u which minimises U, satisfies the differential equation $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$ everywhere in the domain T. This already makes it clear that there always exists a function u having the desired property, namely that function for which U becomes a minimum.
> 2. Every function u which satisfies the [above] differential equation within the domain T, minimises the integral U.
> 3. The integral U can have only one minimum.
>
> It follows from 2 and 3 that there is only one function u with the desired property.

To a physicist this 'principle' seems entirely reasonable. It confirms what nature herself tells you: there is an equilibrium distribution of charge, and so forth. But a mathematician needs to check whether there are any functions at all that extend the function defined on the surface to the whole of the interior or the exterior of the surface. Then it needs to be checked that there are such functions for which the integral is finite; if not, the class of functions Dirichlet wished us to consider is empty and the principle is vacuous. Last, but indeed not least, although Dirichlet said 'it is evident that this integral has a minimum since it cannot become negative', the examples of the function $f(x) = x$ defined on the open interval $(0, 1)$, and of $F(x) = \tanh(x) + 1$ defined on the whole of $\mathbb{R}$, show that a function may be never negative but still fail to have a minimum.

In 1857 Riemann attempted to prove the Dirichlet principle. (He did not simply assume it, as many books say.) He first asserted that if a certain integral over a surface T,

$$L(\alpha, \beta) = \int \left(\left(\frac{\partial \alpha}{\partial x} - \frac{\partial \beta}{\partial y} \right)^2 + \left(\frac{\partial \alpha}{\partial y} + \frac{\partial \beta}{\partial x} \right)^2 \right) dT, \tag{2.6}$$

is finite, where α and β are two arbitrary functions of x and y, then by varying α by a continuous function (or one that is discontinuous only at single points) that is zero everywhere on the boundary of T, the integral attains a minimal value – and moreover, this minimum is attained by a unique function if one excludes the points of discontinuity. He then showed easily that this unique minimising function is harmonic.

Riemann argued that the integral L was defined on a space of functions that was 'a connected domain closed in itself, in which each function can be transformed continuously into every other' (1851, **S**16), from which it followed that there was a sequence of functions in this space for which the

corresponding integral tended to a minimum. Because the space was connected and closed, the limit function was in this space, and so the principle was true.

A charitable interpretation of this remark would be that Riemann expected some theorems to be true because spaces of functions had properties like closed intervals. This is a remarkable insight – but it must also be said that he gave no reason to believe that the space of functions he was considering had the required properties. He did not even define what he meant by the phrase 'closed in itself'.

In any case, this attractive hint was destroyed by one of Riemann's students, Friedrich Prym. He exploited the behaviour of continuous, infinitely oscillating functions to define a continuous function on the boundary of the unit disc that does extend to a harmonic function on the interior of the disc but for which the Dirichlet integral is infinite. This shows that Dirichlet's principle can fail, even when Dirichlet's problem can be solved.

The next person to tackle the Dirichlet problem was Hermann Amandus Schwarz, who developed his solution to this problem in a series of papers around 1870. He first showed that the problem can be solved when the domain T is a disc. He let $f(\phi)$ be an everywhere finite, continuous, real-valued function that is periodic with period 2π – so he could think of f as defined on the unit circle, which is the boundary of the unit disc. He then defined a function $u(r, \phi)$ on the unit disc by the equations

$$u(r,\phi) = \frac{1}{2\pi}\int_0^{2\pi} f(\psi)\frac{1-r^2}{1-2r\cos(\psi-\phi)+r^2}d\psi\,, \quad 0 \le r < 1$$

and

$$u(1,\phi) = f(\phi).$$

It is not entirely obvious that this function is finite and continuous on the whole closed unit disc $0 \le r \le 1$, but Schwarz showed that it is by a careful analysis of the situation as $r \to 1$. Straightforward differentiation then shows that it satisfies the differential equation $\Delta u = 0$ in the interior of the disc, and so the function $f(\phi)$ is a solution of the Dirichlet problem.

It followed from some earlier work of Schwarz's that the problem is therefore also solved for a square. Schwarz therefore proposed to show how the Dirichlet problem could be solved for domains that are composed of an overlapping family of domains, on each of which the problem has already been solved. This is his famous alternating method, which we now outline.

See Schwarz (1870).

Schwarz's alternating method

Schwarz now supposed he had two domains, T_1 and T_2, on each of which separately he had a solution to the Dirichlet problem for any prescribed functions g_1 on the boundary of T_1 and g_2 on the boundary of T_2 (Figure 2.8).

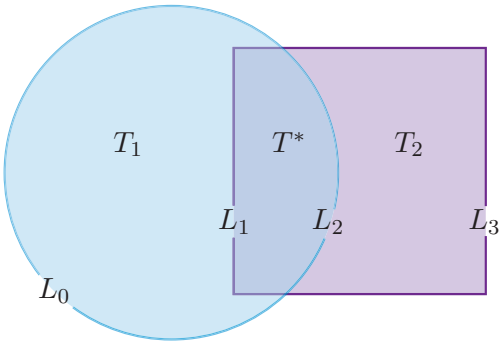

Figure 2.8 A circle and a square

He now set out an ingenious convergence argument. To make it work, he made some simplifying assumptions. First, the regions must overlap nicely; so $T^* = T_1 \cap T_2$, the region common to both of them, is two-dimensional and the boundaries of the regions cross with distinct tangents. Schwarz broke the boundary of each region into two parts: the part of the boundary of T_1 that lies outside T_2 he called L_0; the part of the boundary of T_1 that lies inside T_2 he called L_2; the part of the boundary of T_2 that lies outside T_1 he called L_3; the part of the boundary of T_2 that lies inside T_1 he called L_1.

His aim was to produce a harmonic function on the interior of $T_1 \cup T_2$ that has a continuous extension to a given function g on $L_0 \cup L_3$. Let the minimum value of g on $L_0 \cup L_3$ be k, which without loss of generality may be assumed to be positive.

Schwarz imagined that an air pump pumped air from the region T^* alternately into the two halves of T, the regions $T_1 \setminus T^*$ and $T_2 \setminus T^*$, through the membranes L_1 and L_2, until equilibrium was reached. More formally, he argued as follows.

1. Look at T_1. There is a harmonic function in T_1 that agrees with g on L_0 and takes the constant value k on L_2; call it h_1.
2. Now look at T_2. There is a harmonic function in T_1 that agrees with g on L_3 and with h_1 on L_1; call it h_2. Notice that the values of h_2 on L_1 are greater than or equal to k.
3. Now look again at T_1. There is a harmonic function in T_1 that agrees with g on L_0 and with h_2 on L_2; call it h_3.
4. Now look again at T_2. There is a harmonic function in T_2 that agrees with g on L_3 and with h_3 on L_1; call it h_4.

To conclude the proof, Schwarz showed that the series

$$h' = h_1 + (h_3 - h_1) + (h_5 - h_3) + \cdots + (h_{2i+1} - h_{2i-1}) + \cdots,$$
$$h'' = h_2 + (h_4 - h_2) + (h_6 - h_4) + \cdots + (h_{2i+2} - h_{2i}) + \cdots,$$

converge unconditionally and indeed define a harmonic function on the whole of $T_1 \cup T_2$ that has the prescribed boundary values of the function g on the boundary $L_0 \cup L_3$. So the Dirichlet problem is solved for the union of the two regions, which is what Schwarz set out to do.

Schwarz's proof was a significant step forward, but it remained unclear what class of domains could be treated in this fashion. In particular, could the Dirichlet problem be solved for a region obtained from the overlap of infinitely many regions for which it was solvable? This was the criticism made by Axel Harnack in his book of 1887.

Connectedness – Harnack's work

Harnack considered a sequence of harmonic functions u_n defined on a region F of the plane, where each function u_n restricts to a continuous function U_n on the boundary of F. His idea was to show that as the sum of the functions on the boundary converges to a given function U, the corresponding functions in the interior converge to a harmonic function. He first proved, under suitable conditions on the behaviour of the functions u_n near the boundary, that if the sum $U := \sum_n U_n$ converges uniformly then the sum $u := \sum_n u_n$ converges at every interior point of the surface F to a harmonic function. From this, what became known as Harnack's Theorem followed.

Theorem 2.1 *Harnack's Theorem*

If the members of a sequence of harmonic functions u_n all have the same sign (say, positive) and the sum $u := \Sigma u_n$ converges at an interior point of the surface F, then it converges at every interior point of the surface F to a harmonic function. Alternatively, if the harmonic functions u_n tend from below to the values of a function u, then u is a harmonic function.

Harnack tried to consider regions of a very general kind. He was the first mathematician to try to explain what it is for a domain to be connected. Under his definition, a space is *connected* if any two points can be joined by a finite polygonal arc that can be covered by overlapping discs that all lie in the interior of the domain. We would say that this defines a path-connected space. Later authors were to tease out the concept of a domain as a space in which every point has a disc-like neighbourhood lying entirely in the space.

Harnack defined a *boundary point* of a space as a point such that every neighbourhood of it contains some points that are in the domain and some that are not. He then claimed that the boundary of a simply connected domain is continuous, although it may be nowhere differentiable and may have corners and cusps, and could even have an arbitrary number of incisions into the domain, which would be lines drawn inwards from boundary points that would be traversed twice by any circuit of the boundary.

With these definitions in place, Harnack then proved that the Dirichlet problem can be solved for an arbitrary bounded region of the plane with an arbitrary boundary. He supposed that the general simply connected domain can be steadily approximated by polygonal regions, and used a

limiting argument, which he attributed ultimately to Schwarz, to deal with the general region. He used his theorem to establish that the sequence of harmonic functions on the polygonal regions converges to a harmonic function on the given domain. Finally, he patched together arbitrary bounded domains out of simply connected pieces.

Discs and their boundaries

One consequence of Harnack's work was that it became clear to mathematicians that there is a great variety of curves that can be the boundary of a disc. There is a large class of such boundaries that are in a continuous one-to-one correspondence with a circle. And then there are domains with one or more spikes, such as Figure 2.9 indicates.

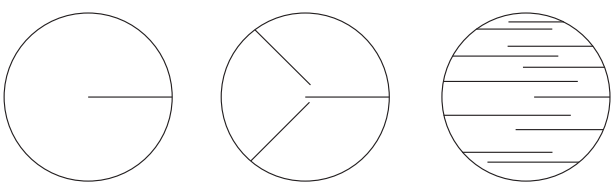

Figure 2.9 Some topological discs

Each of the examples in Figure 2.9 displays a region that is homeomorphic to a disc, but the boundaries are not simple closed curves. One can even contemplate regions with infinitely many spikes. Harnack's work was extended to this setting by the North American mathematician William Fogg Osgood in 1900.

Conversely, one can ask if a region bounded by a simple closed curve is necessarily homeomorphic to a disc. The first person to see that there was something to prove here and to attempt a proof was Camille Jordan, in a note appended to Volume 3 of his *Cours d'analyse* (1887, pp. 587–94). There he outlined a proof, which began by passing from the curve to an arbitrarily close polygonal approximation. Over the years this proof fell into suspicion, and one given by the North American mathematician Oswald Veblen in 1905 came to be regarded as the first correct argument, but a careful recent analysis by Thomas Hales (2007) has shown that Jordan's proof was fundamentally sound and can be polished up into a rigorous argument.

Compactness

The concept of compactness, on the other hand, has a curious history. There are several theorems which say that if a subset X of $\mathbb{R}$ or $\mathbb{R}^n$ is compact then a continuous function on X is uniformly continuous, and is bounded and attains its bounds. For example, Dirichlet in the 1830s proved this claim about continuous functions on a closed interval. The key ingredient in the proof is that any covering of a closed interval by open intervals contains a sub-covering that has only finitely many open intervals. The Heine–Borel Theorem says that a subset of $\mathbb{R}^n$ is compact if, and only if, it is closed and bounded.

The great advantage of compactness is that, at least intuitively, the limit of a sequence of points in a compact set is also in the set. Strictly speaking, the relevant concept here is that of sequential compactness: a space is sequentially compact if every sequence of points in it has a convergent subsequence. Happily for us, if a space is a metric space then it is compact if, and only if, it is sequentially compact.

The advantages of doing analysis on compact sets and compact spaces is clear. Approximations that tend to a limit tend to a limit in that set or space. Moreover, the Bolzano–Weierstrass Theorem says that every sequence on a compact set has a convergent subsequence. Problems in analysis where the solution is found by a sequence of approximations therefore differ markedly depending on whether they are on compact or non-compact spaces, and the sad but predictable news is that analysis mostly happens on non-compact spaces.

Consider, for example, the Dirichlet problem, where we are asked to find a function that minimises a certain integral over a given region. Candidate functions are required to be continuous everywhere inside the region, and agree with a given continuous function on the boundary of the region. One way of thinking of Harnack's Theorem in the present setting is that it tells us that the limit function does not always exist, and that is because the space of candidate functions is not compact.

2.3 Metric spaces – Fréchet

The mathematician who saw first how important compact spaces can be was the French mathematician Maurice Fréchet (Figure 2.10). Indeed, the idea of a metric space was born in his doctoral dissertation, where he discussed the behaviour of limits in what he called 'abstract sets' – he did not speak of spaces or topology until much later.

Figure 2.10 Maurice Fréchet (1878–1973)

In a 1906 paper, Fréchet first defined axiomatically abstract spaces with a concept of limit; then, finding that not much could be said at such an abstract level, he brought in the idea of *separation* ('écart') by adding an axiom that allowed him to talk of the separation between any two elements of an abstract set. He denoted the separation of two elements A and B by (A, B) and imposed the following rules:

$$(A, B) \geq 0,\ (A, B) = 0 \text{ if, } \textit{and only if, } A = B,$$

and if (A, B) and (B, C) are less than some ε then (B, C) is less than some $f(\varepsilon)$, where (ε) is a function that tends to zero as ε tends to zero. He then said that a sequence of elements A_j tends to a limit A if, and only if, $\lim_{j\to\infty}(A_j, A) = 0$. Spaces with a concept of separation Fréchet called *neighbourhood sets* or V-sets (the V standing for 'voisinage' or neighbourhood in French).

At one point in his doctoral thesis, Fréchet introduced the idea of a V-set that obeyed the triangle inequality:

$$(A, B) + (B, C) \geq (A, C).$$

Such a space is a metric space.

Fréchet then developed a number of theorems that are fundamental to point-set topology. For example, he defined a set to be *closed* if every limit point of the set belongs to the set, and *compact* if the intersubsection of every nested sequence of closed sets contains at least one point, and then proved, using methods introduced by Weierstrass, that a continuous function on a compact L-space is bounded and attains its bounds (L-spaces are spaces in which limits exist).

It is perhaps more interesting to note some of the examples Fréchet gave of spaces with limits. These included spaces of functions, spaces of curves or surfaces, and spaces of parameterisations of a given curve, where the notion of convergence is defined by a suitable definition of separation. We illustrate just one of these: the space of all curves in $\mathbb{R}^3$. A curve is defined as the image of three functions $(f(t), g(t), h(t))$ – that is, by a parameterisation, which is moreover defined on the closed unit interval. The separation of two curves C_1 and C_2 with parameterisations $(f_1(t), g_1(t), h_1(t))$ and $(f_2(t), g_2(t), h_2(t))$ is defined as the maximal value of

$$|f_1(t) - f_2(t)| + |g_1(t) - g_2(t)| + |h_1(t) - h_2(t)|,$$

which is finite because the curves lie in a finite region of $\mathbb{R}^3$, and the separation of the curves C_1 and C_2 is defined as the maximum value of their separations under all possible parameterisations.

2.4 Topological spaces – Hausdorff

Felix Hausdorff's *Grundzüge der Mengenlehre* (*Foundations of Set Theory*) of 1914, which he dedicated to Cantor, deals, as its title indicates, extensively with the novel theory of sets before turning to more topological questions in the second half of the book. Hausdorff thought that the greatest success of set theory so far was in clarifying geometrical ideas – where, in his opinion, the new methods were both more profound and simpler than the ones they replaced.

With no more than a mention of Fréchet's name in a footnote, Hausdorff gave a definition of a metric space ('metrische Raum'; the modern term may derive from this work) as one where to each pair of points x, y there is a non-negative number $\overline{xy}$ called their separation or distance ('Entfernung') such that:

- $\overline{xy} = \overline{yx}$
- $\overline{xy} = 0$ if, and only if, $x = y$
- $\overline{xy} + \overline{yz} \geq \overline{xz}$.

He noted that Euclidean n-dimensional space with the usual distance function is a metric space.

Hausdorff then pointed out that in a metric space one can define a neighbourhood of a point x as the set $U_x = \{y \,|\, \overline{xy} < \rho\}$, where ρ is a number greater than zero.

He then defined a topological space as a set with a family of subsets U_x, called the neighbourhoods of x, with the following properties.

1. There is at least one neighbourhood U_x for each x in the space, and every neighbourhood U_x contains x.
2. If U_x and V_x are two neighbourhoods of the same point x then there is a neighbourhood of x contained in $U_x \cap V_x$.
3. If y is a point in U_x then there is a neighbourhood of y that lies entirely inside U_x.
4. Given any two distinct points x and y, there are two neighbourhoods U_x, U_y such that $U_x \cap U_y = \emptyset$.

The last of these axioms now carries Hausdorff's name – it is the Hausdorff Axiom. As he noted, metric spaces are spaces of this kind; later mathematicians dropped this requirement from the definition of a topological space, and in modern mathematics we do not require a topological space to satisfy the Hausdorff Axiom.

Hausdorff now defined many useful topological terms, including interior points of a set, boundary points, closed sets, perfect sets and connected sets. A non-empty set is connected if it cannot be expressed as the union of two disjoint non-empty closed subsets. To define a compact set, he said he followed Fréchet but then gave a different definition: a set in which every infinite subset has an accumulation point (the accumulation point need not be in the set itself). The empty set and all finite sets were also declared to be compact. Under this definition, as Hausdorff noted, every subset of a compact set is compact, and he went on to prove that in Euclidean space (which is not compact even under this definition) a bounded set is compact.

Later in the book he looked at compact metric spaces. He defined a metric space to be bounded if it is contained in a spherical neighbourhood of one of its points, and totally bounded if for every positive number ρ the space is contained in finitely many spherical neighbourhoods of radius ρ. Here Hausdorff observed that a bounded subset of a Euclidean space is totally bounded. He then proved that every compact metric space is totally bounded. Under his definitions, the converse is false. As he noted, the set of rational numbers is totally bounded but not compact.

Our definition of a compact subset of Euclidean space, or any metric space, leads to the theorem that such a subset is compact if, and only if, it is closed and bounded. This leads us to ask when Hausdorff's concept of 'compact' acquired the extra concept of being closed, and so became Fréchet's concept, but a clear historical account seems not to have been written.

In the last thirty years, the study of fractals has made Hausdorff's innovative study of the dimensions of sets into his most cited paper. He published this paper (the title translates as 'Dimension and exterior measure') in 1919. In it, he looked at Euclidean spaces $\mathbb{R}^n$. He supposed that there was a family $\mathcal{U}$ of subsets with the property that for every $\varepsilon > 0$, every subset A of $\mathbb{R}^n$ can be covered by a finite or countable set of sets U

in $\mathcal{U}$, where each U has a diameter of less than ε. He also assigned to each U in $\mathcal{U}$ a non-negative real number $\ell(U)$. Then he defined, for A a subset of $\mathbb{R}^n$ and $\varepsilon > 0$,

$$L_\varepsilon(A) = \inf\{\sum_{j\geq 1} \ell(U_j)\},$$

where the sum is taken over a covering of A by sets in $\mathcal{U}$ of diameter less than ε. He then defined

$$L(A) = \lim_{\varepsilon \to 0} L_\varepsilon(A).$$

This, as Hausdorff showed, defines what is called a positive outer measure on subsets of $\mathbb{R}^n$; that is to say, the following four properties hold:

1. $0 \leq L(A) \leq \infty$, and $L(\emptyset) = 0$
2. $B \subset A$ implies $L(B) \leq L(A)$
3. $A = \bigcup_{j=1}^{\infty} A_j$ implies $L(A) \leq \sum_{j=1}^{\infty} L(A_j)$
4. if sets A and B are a non-zero distance apart then $L(A \cup B) = L(A) \cup L(B)$

along with one too technical to state here, which implied that under certain conditions $L(A)$ was the infimum of all values of $L(B)$ where $B \supset A$.

Then Hausdorff introduced what has become called a Hausdorff function, λ, which is a strictly increasing function from $[0, \infty]$ to $[0, \infty]$ with $\lambda(0) = 0$, and supposed that $\ell(U)$ was of the form $\lambda(d(U))$. This gave him outer measures L^λ, and he said that a subset A of $\mathbb{R}^n$ was of dimension λ if $0 < L^\lambda(A) < \infty$.

Hausdorff was not able to show that there are subsets of $\mathbb{R}^n$ of arbitrary dimension for any choice of Hausdorff function, but he was able to do so if the Hausdorff function was unbounded. His construction was a variant of the construction of the Cantor set. He defined a set of strictly positive numbers $a_j, j = 0, 1, \ldots$ such that $a_j > 2a_{j+1}, j = 0, 1, \ldots$ and $2^j \lambda(a_j) = 1, j = 0, 1, \ldots$. Now he took the interval $[0, a_0]$ and removed the central open interval of length $a_0 - 2a_1$, leaving two closed intervals of length a_1. From each of the remaining two intervals he removed the central intervals of length $a_1 - 2a_2$, leaving four closed intervals of length a_2, and so on. If the union of the 2^j intervals at the end of stage j is called A_j, then Hausdorff considered the set

$$A = \bigcap_{j=1}^{\infty} A_j$$

and showed that it has Hausdorff dimension λ.

It took almost a decade for mathematicians, starting with Besicovitch, to take up these ideas, but their importance has grown ever since. Curiously, this was Hausdorff's only paper on the idea, and he turned instead to work on rather more classical problems in analysis.

3 Biographies

3.1 Pythagoras

One of the earliest, and certainly one of the most famous, mathematicians is said to be Pythagoras, but we know very little about him. It is likely he wrote nothing; certainly no written work ascribed to him survives, and most of the early historical accounts that refer to him make him a religious leader who flourished in Croton in what is now South East Italy around 560 BCE. His followers, the Pythagoreans, adhered to his teaching for some centuries after his death; Aristotle wrote about them some 150 years later, but never directly about Pythagoras himself in any text we have.

The Pythagorean community after his death did contain a number of mathematicians. However, there is actually no evidence that Pythagoras knew, or proved, the theorem that today bears his name, and the famous stories about what happened when the result was discovered, and later revealed, are just that: stories. They say much more about the storytellers, and the needs of their audiences, than they do about the past. In particular, there is no truth at all in the dramatic story that Pythagoras taught that 'All is number', meaning that the discovery of irrational numbers such as the square root of 2 imperilled his entire philosophy and someone was drowned at sea for revealing this awful secret.

The Pythagoreans as a cult continued for many centuries, doubtless with many changes on the way, and there is good evidence that they were absorbed in number mysticism. According to Aristotle, the female was likened by them to the number two, the male to the number three, and their sum, five, was likened to marriage.

What we do know about ancient mathematics is that there is an abundance of evidence that Mesopotamian mathematicians a millennium before had known the Pythagorean theorem as a fact about right-angled triangles, as did mathematicians in other cultures. It is less clear that these cultures had a proof of the result. Somewhere along the line, perhaps in a Greek community, someone must have discovered a proof that $\sqrt{2}$ is irrational, because Aristotle sketched a proof, but the Greeks handled irrational square roots by the simple device of comparing their squares; there is no evidence that they went into a panic about some failure of geometry to reduce to the study of whole numbers and their ratios. By the time of Euclid somebody, most likely a mathematician called Eudoxus, had worked out how to discuss magnitudes of any kind of size; this work forms the impressive Books V and VI of Euclid's *Elements*.

Accurate modern accounts of what we can reasonably say about Pythagoras can be found in the online *Stanford Encyclopedia of Philosophy* (Huffman, 2014) and in the book *The Cult of Pythagoras* (Martinez, 2012).

3.2 Euclid

We may know very little about what Pythagoras did, but we do have scholarly editions of the 13 books that compose Euclid's *Elements*. They give a rich picture of how Greek mathematicians put together the foundations of geometry and then built up a wide variety of results. Because Books VII, VIII and IX are about numbers, and are much more elementary than the books on either side, it is customary to say that in them Euclid collected earlier results (even, some say, Pythagorean ones). More precisely, the first six present a treatise on elementary plane geometry, Book X is on incommensurables and the final books, XI, XII and XIII, turn to the topic of solid geometry and culminate in the construction of the regular solids. Of the three on number theory, Book VII is concerned with various types of number – odd, even, square, prime, composite, etc. – and includes 'Euclid's Algorithm' for finding common factors. Book VIII is concerned with simple properties of squares and cubes, and contains some results on numbers in geometric progression. Book IX contains a number of theorems of special interest, the pearl being Euclid's celebrated proof of the existence of infinitely many primes, which we met in Book A, Chapter 2, Section 3.

Over one thousand editions of Euclid's masterpiece have appeared since the first printed version in 1482 and, to this day, very few books have been circulated and studied more widely. So it would be very pleasant if we could say something about the author of this bestseller, the *Elements*. But very little indeed is known about him. Most histories of mathematics say that he was born around 330 BCE, but even this date is disputed by the leading expert on Greek mathematics, Alexander Jones, who concludes his recent analysis of what Greek and Roman authors said about Euclid as follows (unpublished private communication):

> But the authentic Euclid is still hard to pin down. He does seem to have lived several decades later than we have been accustomed to say, and he was a more advanced and original mathematician than we would assume from the *Elements* and his other surviving works. But beyond this Euclid remains, as he was already in antiquity, an enigma, the most faceless of the great Hellenistic mathematicians.

3.3 Marin Mersenne

Mersenne (Figure 3.1) was a Franciscan friar, spending the main portion of his life in the Minim convent in Paris. From there he corresponded and became acquainted with many of the important mathematicians of that period, including Fermat. Indeed, Mersenne was regarded as a 'centre for correspondence' for mathematics at that time.

In 1644 Mersenne published the work for which he is best known, *Cogitata Physico-Mathematica*. In the preface to this book he made the claim that the only primes $p \leq 257$ for which $M_p = 2^p - 1$ is also prime are 2, 3, 5, 7, 13, 17, 19, 31, 67, 127 and 257. The last of these, 257, is the 55th prime and so this statement includes the claim that $2^p - 1$ is composite for the other 44 primes p that do not exceed 257.

Figure 3.1 Marin Mersenne (1588–1648)

How Mersenne reached his conclusions will never be known, but it is clear that with what was available at that time he could not have tested primality of each of the numbers by hand calculation. In fact it took 303 years and a lot of labour to check out the whole of Mersenne's claim. It is remarkable how close to the truth his claim was, for we now know that he made just five mistakes.

3.4 Pierre de Fermat

The seventeenth century saw three remarkable breakthroughs in mathematics. In the 1630s Descartes successfully introduced algebraic methods into the study of geometry, which made it possible for mathematicians to define and study large classes of curves for the first time. In the 1660s Newton and Leibniz independently invented the calculus. Less successful in its day was the work of Fermat on number theory, but it became one of the most important sources of the subject.

Fermat (Figure 3.2) was born near Toulouse, France, the son of a leather merchant. The year of his birth is uncertain, and he lived a quiet and uneventful life in and around his birthplace; from the age of 30 until shortly before his death in 1665 he was a lawyer and magistrate in the local parliament of Toulouse. His education in advanced mathematics seems to have begun during his time in Bordeaux in the late 1620s, and his first contributions were to the new analytic geometry that was rapidly dominated by Descartes. One offshoot of his work was that he discovered methods for finding tangents to curves and their maxima and minima well before Newton and Leibniz came on the scene.

Figure 3.2 Pierre de Fermat (1601?–1665)

Fermat seldom sought to publish his work and often circulated his ideas among a small group of mostly French mathematicians, and this was the case with his work on number theory. In 1657 he challenged the British mathematicians John Wallis and William Brouncker to show that there are always integer solutions to equations of the form $x^2 - Ay^2 = 1$, where A is positive and has no square factors – a challenge they eventually met – and we know from his choice of specific values of A, such as 109, that Fermat had discovered a general method, because the smallest values of x and y in this case are $x = 158070671986249$ and $y = 15140424455100$. These are not numbers you discover by trial and error.

Fermat's Last Theorem is the subject of Book C, Chapter 12.

One of Fermat's most famous 'theorems' is the 'Last Theorem', which was neither a theorem in his hands nor the last result he left unproved at his death. He owned a copy of Bachet's early seventeenth-century edition of the Greek mathematician Diophantus's *Arithmetica*, and in the margin by Book II Prop. 8, where Diophantus discussed the problem of dividing a given square number into two squares, Fermat wrote:

> On the other hand it is impossible to separate a cube into two cubes, or a biquadrate into two biquadrates, or generally *any power except a square into two powers with the same exponent*. I have discovered a truly marvellous proof of this, which however the margin is not large enough to contain.

The book in which he wrote this marginal note seems to be lost; the note itself was only published after his death, in the edition of his works made by his son Samuel in 1679. Since Fermat wrote nothing else on this subject that has survived, and as no one has ever found a proof with the means available to Fermat (as opposed to those of Andrew Wiles in 1995), the most likely thing is that Fermat simply made a mistake.

See Fermat (1679).

As for Fermat's genuinely last outstanding claim, it concerns the question of which integers can be the areas of right-angled triangles with sides that are rational numbers. Arabic writers knew of a triangle of this kind with area 5, which is now known to be the smallest possible value for the area, but the general problem remains unsolved.

Fermat came to this problem via another problem Diophantus had raised about triangles, in this case ones whose areas plus or minus a number are squares. Bachet, in his edition, remarked that if A is the area of a right-angled triangle with integer sides then there is number K such that $(2A)^2 - K^4$ is a square, and indeed if

$$\left(2pq(p^2 - q^2)\right)^2 + K^4 = s^2$$

then we may set

$$s = 2pq(p^2 + q^2),\ K = 2pq,$$

as you can readily check.

Fermat investigated this in the case when A is required to be a square, and was led to look for integer solutions to $x^4 + y^4$ being a square. He then showed that there is not even a solution in integers to the weaker equation $x^4 + y^4$ equalling a fourth power, which implies that there are no integer solutions to $x^4 + y^4$ equalling a square. In this case we do have Fermat's argument, which shows not only that he did tackle the $n = 4$ case of his 'Last Theorem' successfully but, more importantly, that he introduced into number theory a truly powerful new idea. He argued that one can prove certain equations do not have solutions in positive integers by arguing that *if* they do *then* they have a solution in smaller positive integers. This would give rise to an infinite descending sequence of positive integers, which is impossible. This method, known as Fermat's method of infinite descent, has proved invaluable to the present day.

See Book C, Chapter 12.

3.5 Leonhard Euler

Whilst Fermat's work on number theory had briefly revitalised the subject, he failed to interest many other distinguished mathematicians of his time in his work on numbers. The first to really appreciate Fermat's work was Leonhard Euler (Figure 3.3).

Euler was born in Basel, Switzerland. At the age of 13 he was sent to the University of Basel to study theology so that he might become a clergyman, like his father. However, at university Euler met the son of the distinguished mathematician Johann Bernoulli, who was the leading exponent of Leibniz's calculus, and who invited Euler to visit him regularly

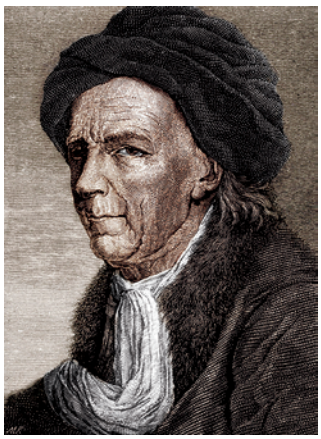

Figure 3.3 Leonhard Euler (1707–1783)

on Saturdays to discuss mathematics. Bernoulli recognised his talents at once, and went on over the years to praise him far more strongly and unreservedly than he ever praised anyone in his own highly talented mathematical family. Finally, in 1745, Bernoulli called Euler 'incomparable, the prince of mathematicians'.

Euler's career took off in 1727 with a call to the new Academy being set up in St Petersburg – the city Tsar Peter the Great had created as the face of modern Russia, in many ways the opposite of Moscow. The Tsar had decided to found an Academy of Science there in 1724, and when he died in 1725 his widow Catherine I took it over. Johann Bernoulli declined an invitation to move there, but his sons Niklaus II and Daniel I did go, and they and Christian Goldbach, the first permanent secretary of the Academy, lobbied for a position for Euler.

Euler's time there was generally a happy and productive one. His output was extraordinary. He had published his first paper (on a design for the arrangement of masts on a ship) at the age of 18, before going to St Petersburg – and before he had ever seen a ship. He went on to publish over 500 papers and many books during his lifetime. After his death, it took the St Petersburg Academy 47 years to publish all the manuscripts he had left behind. His writings covered the complete spectrum of mathematics including applications of calculus, algebra, coordinate and differential geometry, infinite series and number theory. He survived a life-threatening illness in 1735, but another illness that flared up in 1738 cost him the sight of his right eye. (He did not lose his sight through overwork, as is sometimes asserted.)

In 1740 Euler decided to leave St Petersburg for Berlin, to manage the Academy (later, the Académie Royale des Sciences et des Belles-Lettres) that the Prussian Emperor Frederick the Great was establishing in Berlin with Maupertuis as its first President. He stayed until he came to realise that he had accomplished all he could there, leaving Berlin on 1 June 1766 and returning to St Petersburg on 28 July, to a triumphant reception. There he stayed, productive as ever even though the sight in his remaining eye was deteriorating, until on the afternoon of 18 September 1783 he was felled by a massive stroke. In the words of the French mathematician the Marquis de Condorcet, at around 11 o'clock that night 'The great Euler ceased to calculate and to live'.

Euler's *Collected Works* runs to 78 volumes, and undoubtedly one of the results of his prolific publication at many levels was to make mathematics accessible to a wider audience than ever before. Four of these volumes record his explorations of number theory, and more results are to be found elsewhere. His extensive calculations with binary quadratic forms came close to discovering the law of quadratic reciprocity. The zeta function is one of his discoveries, as is the remarkable result that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

Euler brought the subject out of the backwater to which generations had consigned Fermat's work and made it worthy of the attentions of Lagrange and, in due course, Gauss.

3.6 Joseph-Louis Lagrange

Lagrange (Figure 3.4) was born in Turin to a French father and Italian mother, but he spent the most productive years of his life in Berlin and then Paris. His father intended him to become a lawyer, but Lagrange and others soon realised that he had a particular talent for mathematics, and he became appointed a professor at the Royal Artillery School in Turin at the age of 19.

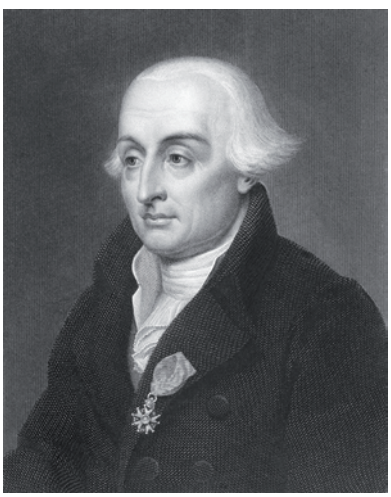

Figure 3.4 Joseph-Louis Lagrange (1736–1813)

Lagrange sent many of his earliest discoveries to Euler, who at the time worked at the Academy of Sciences in Berlin, and although Euler recognised a great talent in the young Lagrange they never met. In 1766, when he was due to leave the Berlin Academy for St Petersburg, Euler arranged for Lagrange to be installed as his replacement. A message from Frederick the Great invited 'the greatest mathematician in Europe' to live near 'the greatest king in Europe'. Lagrange accepted, and worked in Berlin for 20 years until, on the death of Frederick in 1786, he was head-hunted by the French and moved to Paris the next year.

For the next six years Lagrange contributed nothing to mathematics. Strangely, it was the turmoil of the French Revolution that revitalised his interest in the subject. Following the abolition of the old universities, Lagrange became a professor at the newly founded École Polytechnique. In this period he published his attempt to establish the calculus on a purely algebraic footing, the *Théorie des fonctions analytiques* (1797).

Although Lagrange's main researches were in the domains of analytical mechanics, theoretical astronomy, calculus and function theory, he was also a profound number theorist. In 1770 he was the first to prove Fermat's claim that every positive integer can be expressed as the sum of four squares, and he often turned Euler's provocative speculations into proven theorems. He was the first to propose that a binary quadratic form $ax^2 + bxy + cy^2$ can be reduced to a form in a finite list of such forms with the same value of $b^2 - 4ac$; Lagrange's reduction theory was a significant step in bringing coherence to the findings of Fermat and Euler, and Euler was particularly pleased by it.

3.7 Adrien-Marie Legendre

Legendre (Figure 3.5) was born in Toulouse and educated at the Collège Mazarin in Paris. Throughout a distinguished mathematical career he contributed to significant advances in many fields. He was mainly an analyst, who wrote a three-volume comprehensive treatment of integral calculus, followed by a further three-volume work on elliptic integrals. In these works he introduced numerous tools of analysis still used by applied mathematicians, including the beta and gamma functions and the

Legendre polynomials, which are used in solving certain differential equations. He also wrote an introduction to geometry, his *Éléments de géométrie*, that was influential in swinging geometry teaching back to the formal style of Euclid and remained in use in France for much of the nineteenth century, but which was marred in successive editions by flawed attempts at defending the parallel postulate and therefore denying the possibility of non-Euclidean geometry.

Figure 3.5 Adrien-Marie Legendre (1752–1833), shown on the left with Fourier beside him; this splendid cartoon seems to be the only authentic picture of Legendre in existence!

Legendre was not a great innovator in number theory but his writings did point mathematicians of the nineteenth century towards fruitful areas for investigation. In 1798 he published a two-volume work, *Essai sur la théorie des nombres*, probably the first treatise devoted solely to number theory; an expanded edition, *Théorie des nombres*, was printed in 1830. Amongst many other results, he completed an argument begun by the young Dirichlet in 1825 that proved the equation $x^5 + y^5 = z^5$ has no solution in the natural numbers – this is a particular case of Fermat's Last Theorem. On the question of distribution of primes, he conjectured that $\pi(n)$ approaches $\dfrac{n}{\log(n) - 1.08366}$ as n increases, a result that is very close to being correct.

In his *Essai* Legendre included an account, together with an attempted proof, of the Law of Quadratic Reciprocity. The relationship given in the Law was first conjectured by Euler in 1783, on the basis of considerable numerical evidence. Two years later Legendre offered the first proof. This proof, like the later one published in the *Essai*, was seriously flawed. It broke into eight cases, in one of which Legendre relied on the plausible

claim that every arithmetic progression $\{a + bk : k = 1, 2, \ldots\}$ with a and b relatively prime contains infinitely many primes. However, he could not prove this claim. Worse, in another of the eight cases Legendre's reasoning was circular. Gauss, in his *Disquisitiones Arithmeticae* (1801), acknowledged that, before him, Euler had almost discovered the result and that Legendre knew it and had proved it in various special cases, but he attributed the Law to himself on the grounds that his two proofs were the first valid ones. This Legendre was unwilling to accept, but the result about primes in arithmetic progressions was only to be established by Dirichlet in 1837, and the circularity in Legendre's line of reasoning has not been removed to this day.

3.8 Joseph Fourier

Joseph Fourier (Figure 3.6) was born in Auxerre in France and orphaned at the age of 9, at which point he was placed in the town's military school. There he discovered both a passion for mathematics and the strength of his calling to civic and military duty. The French Revolution found him back in his home town as a school teacher. He became involved in politics and courageously opposed the Terror, the bloodthirsty events of 1794 that saw many people dragged without trial to their deaths on the guillotine, and he was arrested. A personal appeal to Robespierre failed, but when Robespierre was himself executed on 28 July 1794, Fourier was released and went to the École Normale. After a year he was appointed as an assistant lecturer at the École Polytechnique, working under Lagrange and Monge, the French geometer who was the founding father of the École Polytechnique. In 1798 Monge, by now a prominent supporter of Napoleon, selected him to go on the French expedition to Egypt. This was a French attempt to exclude the British from the eastern Mediterranean, but it had a colonial, anthropological side, and French engineers made impressive surveys of Egyptian pyramids and other remains.

Figure 3.6 Joseph Fourier (1768–1830)

With the defeat of the French by the British in 1801, Fourier returned to France, hoping to resume his work at the École Polytechnique, but Napoleon had noticed Fourier's impressive organisational talents and sent him instead to be the Prefect (or Governor) of the Department of Isère, which at that time extended from Grenoble to the French border. He did this so well that Napoleon made him a baron in 1808. In 1809 he finished his contribution to the *Description d'Egypt*, a massive account of ancient Egypt based on the 1798 surveys, which marks the start of the celebration of ancient Egypt in the modern world. (Cleopatra's and other needles began their journeys at this time, as did more and more fanciful statements about ancient wisdom, secret knowledge and so forth, none of which can be held against Fourier.)

The eventual defeat and final exile of Napoleon were the lowest point of Fourier's life, but in 1816 a former colleague on the Egypt expedition rescued him and gave him a job as Director of the Bureau of Statistics for the Department of the Seine, a position that left him good time for research. He now suffered for his former associations with Napoleon, but

after more protests, he was appointed to the reformed Academy of Sciences, rising to become its permanent secretary in 1822. He was elected to the prestigious Académie Française in 1827 (by which time he was 59). He died in 1830 as the result of complications from an illness caught in Egypt.

Fourier's great work on heat diffusion, the *Théorie analytique de la chaleur*, has been hailed as one of the few major contributions to both pure and applied mathematics. It was written in several stages. The first version was finished in 1807 and submitted to the Paris Academy of Sciences, but Lagrange blocked publication because Fourier's ideas contradicted some that Lagrange had published in the 1750s. In 1810 the Academy of Sciences announced a prize competition on heat diffusion. Fourier revised his memoir and won the competition, but the memoir was still criticised for a lack of rigour and generality. Fourier thought the criticism unfair, but revised it for publication, and the resulting book came out in 1822, when his personal circumstances had much improved.

3.9 Carl Friedrich Gauss

Figure 3.7 Carl Friedrich Gauss (1777–1855)

Gauss (Figure 3.7) was unquestionably one of the greatest mathematicians of all time, and it is clear from his work that he possessed an ability, rare among mathematicians, for doing elaborate arithmetical calculations in his head. He himself told the story of being a young child in school and surprising his teacher by summing an arithmetic progression immediately the question was posed. But if that story survives in a reputable source, many of the details in histories of mathematics have been added for effect, such as the brutality of the teacher and the actual arithmetic progression (commonly said to be $1 + 2 + \cdots + 100$).

However, we do not need to exaggerate Gauss's merits. He made his name in 1801, the year he turned 24, for two remarkable achievements. The first was in astronomy. On 1 January 1801 the Italian astronomer G. Piazzi had discovered what we now call the asteroid Ceres, the first new body in the solar system since Herschel's discovery of Uranus 20 years before. But it was already close to the sun, and Piazzi was only able to observe it for 42 days before it vanished behind the sun; thus his observations, with all the observational errors they necessarily contained, did not make it easy to determine its orbit. Piazzi published his data in June, and invited Europe's astronomers to predict where it would next be visible. Gauss took up the challenge, and although his prediction was quite widely separated from the others it was clear by the end of 1801 that he was correct. He had succeeded where the others failed because he had been willing to take on two demanding tasks. He allowed that Ceres might have an elliptical orbit rather than a circular one, and he invented the method of least squares to handle Piazzi's data. Such was the excitement caused by the new 'planet' that Gauss briefly tasted European fame – and also controversy, because he did not reveal his methods, so when the distinguished French mathematician Legendre discovered the method of least squares a few years later and published it, he refused to accept Gauss's claims to priority.

Gauss's second great discovery, and the one of more interest to us, was his reformulation of number theory, which he presented in his book *Disquisitiones Arithmeticae* (*Arithmetical Researches*) in 1801. It opens with an account of modular arithmetic, in which he was the first to prove that the multiplicative group of non-zero residues modulo a prime p is cyclic; a generator is called a primitive root mod p. He then showed how modular arithmetic enables one to take up the theory of quadratic forms and solve most of the outstanding problems in the subject. Since this was a topic that Euler, Lagrange and Legendre had repeatedly tackled, Gauss was aiming high indeed.

Some of the better-known results by Gauss's day were as follows.

1. Every odd number of the form $x^2 + y^2$ is a product of prime numbers of the form $4n + 1$ and squares.
2. Every prime number of the form $4n + 1$ is a sum of two squares (i.e. is of the form $x^2 + y^2$).
3. Every divisor of a number of the form $x^2 + 2y^2$ is of that form, and every divisor of a number of the form $x^2 + 3y^2$ is again of that form ...
4. ... but divisors of numbers of the form $x^2 + 5y^2$ are either of that form or of the form $2x^2 + 2xy + 3y^2$ (consider $1^2 + 5 \times 2^2 = 21 = 3 \times 7$).
5. For a given quadratic form $ax^2 + 2bxy + cy^2$, where a, b, c have no common factors, and a given integer m, if there are integers x_0 and y_0 such that $ax_0^2 + 2bx_0y_0 + cy_0^2 = m$ then $b^2 - ac$ is a square mod m. (This is why modular arithmetic enters the theory.) We say that in this case the quadratic form represents the number m.

Gauss took the subject and gave a series of tests that one could apply to find the possible quadratic forms that represent a given number. His tests are easy to apply, his analysis greatly extends and deepens what was known, and his proofs are extremely difficult. This is easy for us to see. In essence he found a way of applying ideas from group theory to the study of quadratic forms, and not just group theory but the study of group characters (maps from a group to the group of non-zero complex numbers). Indeed, the word 'character' in this context was introduced by him. But – and it is a very important 'but' – group theory did not exist in his day, and was not to do so for a generation!

Among the other important results in the *Disquisitiones Arithmeticae* are the first two proofs of the Law of Quadratic Reciprocity, which makes it very easy to handle the question of when one prime is a square modulo another. This occasioned another dispute with Legendre, who had previously claimed a proof of this Law, but one that Gauss had to observe was seriously flawed.

The subsection of the *Disquisitiones Arithmeticae* that proved most popular goes by the unwieldy name of cyclotomy, or circle division. It concerns what can be said about the roots of unity. Because

$$x^n - 1 = (x - 1)(x^{n-1} + x^{n-2} + \cdots + x + 1),$$

the roots of the equation $x^n - 1 = 0$ other than 1 satisfy the equation

$$x^{n-1} + x^{n-2} + \cdots + x + 1 = 0.$$

Even before going to university, Gauss had discovered that this equation can usually be solved by solving a sequence of equations of lower degree. For example, when $n = 17$, the equation for the non-trivial 17th roots of unity is

$$x^{16} + x^{15} + \cdots + x + 1 = 0.$$

Gauss showed that it is possible to find a quadratic equation the roots of which allow one to factorise this equation into two factors of degree 8, and then to repeat this trick by solving another quadratic, the roots of which allow one to factorise each of these polynomial factors of degree 8 into two factors of degree 4, and so on. In the end all the roots of unity are found by solving nothing worse than quadratics, and because solving a quadratic equation is equivalent to carrying out a construction by straightedge and circle, Gauss had found the first regular polygon to be constructible that way since those known to the Greeks. The Greeks had given constructions for the regular or equilateral triangle, the square, the regular pentagon, and all the regular figures constructible from these by bisecting their sides, but the regular 17-gon was new. Gauss went on to claim, without giving a proof, that a regular N-gon can be constructed with compass and straightedge if, and only if, $N = 2^r p_1 p_2 \cdots p_k$, where the primes p_i are distinct Fermat numbers and $r \geq 0$. (A Fermat prime is a prime number of the form $2^{2^n} + 1$, such as 3, 5 and 17.)

Gauss retained a lifelong interest in astronomy, and did a lot of work on the orbits of the first few asteroids to be discovered. He was employed in the 1820s to conduct a survey of the state of Hanover, where he lived (the Hanoverian monarchs of England are cousins of the German Hanoverians), and this led him to invent the heliotrope and to discover the concept of the intrinsic curvature of a surface, which decisively reformulated the subject of differential geometry. In the 1830s he did important work on potential theory, which he applied to electricity and magnetism. He also had original things to say about differential equations and complex analysis, and left many profound ideas unpublished at his death. But it seems that number theory was his favourite subject, because of the contrast between the ease with which results can be conjectured and the frequent difficulty found in proving them. Moreover, the proofs often made unexpected and deep connections with other branches of mathematics. To that end, for example, he published a total of six proofs of quadratic reciprocity, and left a further two unpublished.

As a young man, Gauss kept a diary in which to record his discoveries. It did not come to light until some forty years after his death, and its brief statements of 146 results amply display his genius. The first entry concerns cyclotomy, and runs ‘The principles upon which the division of the circle depend, and geometrical divisibility of the same into seventeen parts, etc.’ (30 March 1796, when Gauss was still 18); the sixteenth (for 27 June 1796) concerns the second proof of the Law of Quadratic

Reciprocity and says 'A new proof of the golden theorem all at once, from scratch, different, and not a little elegant'.

3.10 Peter Gustav Lejeune Dirichlet

Dirichlet (Figure 3.8) is one of the most important mathematicians of the nineteenth century. A German despite the French appearance of his name, he was born in Düren in 1805. At the age of 17 he took himself off to Paris to study mathematics, for the level in Germany was very low and Paris was then by far the leading place for mathematics in the world. He made many good social contacts; through them he met Fourier, and acquired from him a strong interest in trigonometric series and mathematical physics. His own interests in mathematics, however, initially concentrated on number theory. At the age of 20 he presented a memoir to the French Academy of Sciences that almost completely resolved the first new case of Fermat's Last Theorem since Euler had dealt with cubes, by showing that many cases of the equation $x^5 + y^5 = z^5$ cannot be solved in integers. The remaining cases were solved soon afterwards by Legendre, but Legendre's publication unduly minimised the extent to which the credit was due to Dirichlet.

Figure 3.8 Peter Gustav Lejeune Dirichlet (1805–1859)

Dirichlet returned to Germany in 1826, and in 1828 was appointed a Professor of Mathematics at the new University of Berlin, founded by Alexander von Humboldt in 1809. In 1831 he married Rebecca Mendelssohn-Bartholdy, the sister of the famous composer; Dirichlet was a good pianist in his own right. It would take us too far afield to discuss Dirichlet's many contributions to number theory, but he became the person who explained Gauss's work to the rest of the world, thus helping to make number theory into a major branch of pure mathematics, and almost the leading branch in Germany.

Equally important was Dirichlet's attitude to mathematical rigour, captured marvellously in the following remark of his lifelong friend Carl Jacobi:

> If Gauss says he has proved something, it seems very probable to me; if Cauchy says so, it is about as likely as not; if Dirichlet says so, it is certain. I would gladly not get involved in such delicacies.

Jacobi to A. von Humboldt, 1846, quoted in H. Pieper's 'Gegen die Schmach des Belagerungszustands' (1980, p. 23). The title means 'Against the ignominy of a state of emergency'.

Dirichlet may be called the man who brought rigorous mathematics to Germany, if not indeed to the mathematical world, and he is particularly remembered for his work on potential theory and his work on the convergence of Fourier series, which could serve as a textbook example of how to do epsilon-delta analysis to this day.

When Gauss died in 1855 Dirichlet, who was then 50, moved to Göttingen as his successor. But his hopes of a new and active phase of life were cut short by a heart attack when he was 58, and within a year he was dead.

3.11 Bernard Riemann

Riemann (Figure 3.9) was the archetype of the shy mathematician, close to his family and awkward outside it. We learn from a recent biography of him that his interests were almost exclusively mathematics, physics and philosophy. He was devout in his religion, and had initially intended to study theology at Göttingen University, in accordance with his father's wishes, before his remarkable ability at mathematics led him to switch subjects. But even though Gauss was a professor there, the level of mathematical education in Göttingen was not particularly high, and in accordance with the German freedom to study anywhere Riemann left Göttingen for Berlin in 1847, where he spent two years learning from Dirichlet about potential theory and partial differential equations, number theory and theory of integration.

See Laugwitz (2000).

Figure 3.9 Bernard Riemann (1826–1866)

But Berlin was a bustling capital, and in other ways Göttingen was the ideal choice for Riemann. He returned there in 1849 and attended Wilhelm Weber's lectures on mathematical physics, and for a while he devoted himself to studies in physics and philosophy of science. This influenced his way of dealing with geometry. In Göttingen he soon met Richard Dedekind, who was five years younger than him, and who was also a truly conceptual thinker, but unlike Riemann one who was to move slowly and steadily throughout a long life.

In 1855 matters improved greatly for Riemann. Gauss had for some years been a distant but real presence in Göttingen, but after his death in 1855 Dirichlet was called as the senior professor. Dirichlet was the ideal abstract thinker to guide Riemann, and the young man remained very grateful to him. His influence is visible in Riemann's work, not just in the published papers but in the lectures he began to give.

When Dirichlet died in 1859, Riemann was quickly appointed his successor (the more charismatic figure of Dedekind being by then a professor in Zürich). He still attracted only a few students, and one wonders how matters might have developed had he not contracted pleurisy in 1862. This led to a permanent weakening of his lungs, and he was advised to spend as much time as possible in the South. He accordingly spent as much time as he could in Italy, but his health deteriorated and on 20 July 1866 he died near Lake Maggiore, where he is buried. He left behind a number of published papers, several more in a good enough state to be published, and yet more that could be edited and printed in the first edition of his *Werke* (edited by Dedekind and Heinrich Weber in 1876).

See Riemann (1851, 1854).

3.12 Georg Cantor

Cantor (Figure 3.10) was born in St Petersburg, Russia, but his family soon moved to Germany. He studied mathematics in Berlin, where he was greatly influenced by Weierstrass, Kummer and Kronecker, the triumvirate who dominated mathematics there in the second half of the nineteenth century. After writing his doctorate he obtained a teaching position at the University of Halle, where he remained for the rest of his life, becoming a

full Professor there in 1879, although he always hoped to be offered a position back in Berlin.

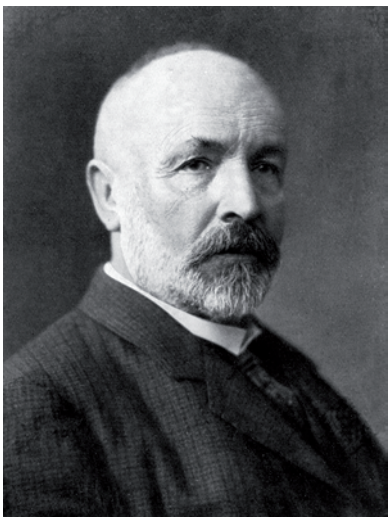

Figure 3.10 Georg Cantor (1845–1918)

Cantor's first work was on the anomalous behaviour of Fourier series and this led him to develop important ideas in point-set topology. Only several years later did he realise that this work opened the door to a new theory of infinite sets in which it was possible to construct sets of different orders of infinity. In 1873 he gave the first proof that the real numbers are not countable, and thereafter his research was focused on the task of creating a rigorous theory of transfinite sets. Opinions of this work varied. Weierstrass was impressed by it, and for a time the Swedish mathematician and energetic journal editor Gösta Mittag-Leffler published it in his journal *Acta Mathematica*; when Cantor thought Mittag-Leffler was losing interest in it he switched to another recently founded journal, Felix Klein's *Mathematische Annalen*. But others opposed it, notably Kronecker, and the big Berlin-based mathematical journal, Crelle's *Journal für Mathematik*, was closed to him. Hermite, the influential French mathematician, found Cantor's work pointless, but French mathematicians of the next generation – such as Poincaré – found the topology useful.

By 1900, Cantor's work had become caught up in heated debates about the foundations of mathematics. It proved easy to find paradoxes in the naive theory of sets, and although Cantor regarded these as harmless and inevitable when finite human minds deal with the concept of infinity, others realised that it was perhaps better to build up a less naive but more rigorous account of set theory. As a result, this became a task for mathematicians in the twentieth century. David Hilbert made Cantor's continuum hypothesis the very first of the problems he posed the international community in 1900. It addresses the size of the set of all real numbers, which is denoted c. The smallest infinite set is the integers, $\mathbb{Z}$, and it has a size denoted $\aleph_0$. Cantor proved that the set of all subsets of $\mathbb{Z}$ is strictly bigger than $\mathbb{Z}$ and said it has size $\aleph_1$. The set of all subsets of this set is bigger again and is said to have size $\aleph_2$, and so on. The question arises of where c fits in to all of this, and Cantor's hypothesis was that $c = \aleph_1$. He spent many years trying to prove it, unsuccessfully; it is now known to be independent of the usual axioms for set theory that were developed after Cantor's death.

Cantor was the first President of the German Mathematical Society when it was established in 1890, but he began to suffer prolonged disabling bouts of depression and was frequently hospitalised. He eventually died in the psychiatric clinic of Halle University.

3.13 Felix Hausdorff

Hausdorff (Figure 3.11) was born in Breslau, Germany (now Wroclaw, Poland) in 1868. He became attracted to philosophy when a student; under the name of Paul Mongré, he went on to write two books on philosophy in an upbeat, Nietzschean spirit, as well as a successful play, before turning to mathematics in his forties. His philosophy led him to wonder what, if anything, could be provably true about the world we live in, and he decided

Figure 3.11 Felix Hausdorff (1868–1942)

that the ideas that time is linear, continuous, single-valued, does not run backwards and proceeds uniformly, and that space is three-dimensional, continuous, uniform and Euclidean, are incapable of proof.

To support his views Hausdorff turned to mathematics, especially Cantor's new theory of infinite sets, and after 1911 his interests became almost exclusively mathematical. His *Grundzüge der Mengenlehre* (1914) marked a significant advance in both set theory and topology. In 1915 Hausdorff published what became known as his paradox, which shows that on any definition of measure that satisfies Lebesgue's four axioms there will be non-measurable sets.

For many years after the First World War he taught mathematics at the University of Bonn, but the Nazi accession to power in 1933 led to years of persecution, because Hausdorff was a Jew. With his wife, he committed suicide in the town of Endenich in January 1942 to avoid deportation to a concentration camp, bravely writing the note 'Und vielleicht ist Endenich auch das Ende nicht' ('and perhaps Endenich is also not the end').

References

Cardano, G. (1663) 'Vita Ludovici Ferrarii Bononiensis', in *Opera Omnia*, vol. 9, Lyons.

Cardano, G. (1543) *Ars Magna* (trans. T.R. Witmer), MIT Press (this edition 1968).

Cauchy, A.L. (1815) 'Mémoire sur le nombre des valeurs qu'une fonction peut acquérir, [etc.]', *Journal de l'École Polytechnique*, vol. 10, pp. 1–27, in *Oeuvres Complètes*, series 2, vol. 1, pp. 64–90.

Cauchy, A.L. (1844) 'Mémoire sur les arrangements que l'on peut former avec des letters données, [etc.]', *Exercises d'analyse et de physique mathématique*, vol. 3, pp. 151–252, in *Oeuvres Complètes*, series 2, vol. 13, pp. 171–282.

Dirichlet, P.G.L. (1876) *Vorlesungen über die im ungekehrten Verhältniss des Quadrats der Entfernung wirkenden Kräfte*, Leibzig, Teubner.

Fermat, P. (1679) *Varia Opera Mathematica*, Toulouse.

Fourier, J. (1822) *Théorie analytique de la chaleur*, Paris, Firmin Didot; English translation *The Analytical Theory of Heat*, Cambridge University Press, 1878 (reprinted Dover and New York, 1950).

Fréchet, M. (1906) 'Sur quelques points du calcul fonctionnel', *Rendiconti del Circolo Matematico di Palermo*, vol. 22, pp. 1–74.

Gauss, C.F. (1801) *Disquisitiones Arithmeticae* (trans. W.C. Waterhouse and A.A. Clarke), New York, Springer (revised edition 1986).

Hales, T.C. (2007) 'Jordan's proof of the Jordan curve theorem', *Studies in Logic, Grammar and Rhetoric*, vol. 10, no. 23, pp. 45–60.

Harnack, A. (1887) *Die Grundlagen der Theorie des logarithmischen Potentiales und der eindeutigen Potentialfunktion in der Ebene*, Liepzig, Teubner.

Hausdorff, F. (1914) *Grundzüge der Mengenlehre*, Leipzig, Veit.

Huffman, C. (2014) 'Pythagoras', in Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy* [Online], Spring 2014 edn. Available at http://plato.stanford.edu/archives/spr2014/entries/pythagoras (Accessed 2 May 2014).

Jordan, C. (1870) *Traité des substitutions et des équations algébriques*, Paris, Gauthier-Villars.

Jordan, C. (1887) *Cours d'analyse de l'École Polytechnique*, vol. 3, Paris, Gauthier–Villars.

Lützen, J. (2009) 'Why was Wantzel overlooked for a century? The changing importance of an impossibility result', *Historia Mathematica*, vol. 36, pp. 374–94.

Lagrange, J.-L. (1797) *Théorie des fonctions analytiques*, Paris, Imprimerie de la République; 2nd edn Paris, Courcier, 1813 in *Oeuvres Complètes*, vol. 9.

Lagrange, J.-L. (1770) 'Réflexions sur la résolution algébrique des équations', *Nouveaux mémoires de l'Académie royale des sciences et belles-lettres de Berlin*, in *Oeuvres Complètes*, vol. 3, pp. 205–421.

Laugwitz, D. (2000) *Brenhard Riemann*, Berlin, Birkhäuser.

Legendre, A.-M. (1798) *Essai sur la théorie des nombres*, Paris, Courcier; 2nd edn 1808.

Legendre, A.-M. (1830) *Théorie des nombres*, 3rd edn, Paris, Firmin-Didot.

Martinez, A.A. (2012) *The Cult of Pythagoras: Math and Myths*, University of Pittsburgh Press.

Morley, H. (1854) *The Life of Girolamo Cardano, of Milan, Physician*, vol. 1, London, Chapman and Hall.

Neumann, P.M. (2011) *The Mathematical Writings of Évariste Galois*, Heritage of European Mathematics series, Zürich, European Mathematical Society.

Pieper, H. (1980) 'Gegen die Schmach des Belagerungszustands', *Spectrum, Monatsbericht der AdW der DDR*, vol. 1, pp. 22–4.

Riemann, B. (1851) 'Grundlagen für eine allgemeine Theorie der Functionen einer veränderlichen complexen Grösse', in *Werke*, pp. 35–80; this edition 1990, ed. R. Narasimhan, Springer.

Riemann, B. (1854) 'Über die Darstellbarkeit einer Function durch einer trigonometrische Reihe', *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen*, vol. 13, pp. 87–132, in *Werke*, pp. 227–71; this edition 1990, ed. R. Narasimhan, Springer.

Robson, E. (2007) 'Mesopotamian mathematics', in Katz, V. (ed.) *The Mathematics of Egypt, Mesopotamia, China, India, and Islam: a Sourcebook*, Princeton University Press, pp. 58–186.

Rothman, T. (1989) *Science à la Mode*, Princeton University Press.

Schwarz, H.A. (1870) 'Zur integration der partiellen Differentialgleichung $\Delta u = 0$', *Monatsberichte der Königlich preussischen Akademie der Wissenschaften zu Berlin*, pp. 767–95, in *Gesammelte mathematische Abhandlungen*, vol. 2, pp. 144–71.

Serret, J.-A. (1849) *Cours d'algèbre supèrieure*, 1st edn, Paris, Gauthier-Villars; 2nd edn 1854, 3rd edn 1866.

Smith, H.J.S. (1875) 'On the integration of discontinuous functions', *Proceedings of the London Mathematical Society*, vol. 6, pp. 140–53, in *Collected Mathematical Papers*, vol. 1, pp. 86–100, Oxford University Press (this edition 1965).

Steinitz, E. (1910) 'Algebraische Theorie der Körper', *Journal für Mathematik*, vol. 137, pp. 167–309; reprinted 1930 with additional material by R. Baer and H. Hasse, Leipzig, de Gruyter.

Tignol, J.-P. (2001) *Galois' Theory of Algebraic Equations*, World Scientific.

Veblen, O. (1905) 'Theory on plane curves in non-metrical analysis situs', *Transactions of the American Mathematical Society*, vol. 6, pp. 83–98.

Wantzel, P.L. (1837) 'Recherches sur les moyens de reconnaitre si un problème de géométrie peut se résoudre avec la règle et le compas', *Journal des mathématiques pures et appliquées*, vol. 2, pp. 366–72.

Acknowledgements

Grateful acknowledgement is made to the following sources.

Figure 1.1: Wellcome Library, London. This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/4.0/.

Figure 1.2: Municipal Archives of Trondheim/Wikimedia.org. This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/2.0/.

Figure 1.3: Wellcome Library, London. This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/4.0/.

Figure 2.7: Pruneau/Wikimedia.org. This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/.

Figure 2.10: MacTutor History of Mathematics archive.

Figure 3.2: © Bettmann/CORBIS.

Figure 3.3: © INTERFOTO/Alamy.

Figure 3.4: © iStock.com/GeorgiosArt.

Figure 3.5: Licensed with permission from Bibliothque de L'Institut de France/RMN-Grand Palais.

Figure 3.6: Stock Montage/Getty Images.

Figure 3.7: © Heritage Image Partnership Ltd/Alamy.

Figure 3.10: © INTERFOTO/Alamy.

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Accessibility for Mathematics and Statistics modules

Printable page generated Tuesday, 13 Aug 2019, 10:52

Accessibility for Mathematics and Statistics modules

Introduction

The Open University provides a range of support services for disabled students. You may not think of yourself as disabled but if you have a long term health condition, specific learning difficulty (such as dyslexia) or mental health difficulty then you are entitled to support. We use the single term ‘disabled’ or ‘disability’ to cover all these conditions. You can find more information about the services that are available by going to the Supporting students with disabilities website which links through to Overcoming Barriers to study if you have a disability or health condition. For example, you may need to have module materials in an alternative format, extra tutorial support or flexibility with some assignment deadlines because of your disability or condition.

If you live in the UK, you may be able to get a grant called the Disabled Students’ Allowance (DSA). This aims to cover the extra costs of studying in higher education – for example, the need for specialist equipment that might arise because of your disability, medical condition or specific learning difficulty. Further information on the DSA is available on the who provides support and adjustments website, and you can also visit Disabled student support. If you would like to discuss your needs, before registering as a student, please contact us. Any sensitive information will be treated confidentially.

If you are already an Open University student, you can also find more information on the Disabled student support website. Once you have read this guide, if you think that you will need additional support, you can submit a Disability Support Form outlining your requirements and then discuss the help that you need with an adviser; if you will need help to complete the form over the phone, please contact us.

You can also obtain advice and information from your Student Support Team.

> Please contact us as soon as possible to discuss your requirements, as some support services can take several months to arrange.

This guide outlines some of the general accessibility issues that apply to distance learning of mathematics and statistics. Many of the issues discussed apply to both mathematics and statistics, and you can assume that references to mathematics also apply to statistics. This guide includes advice on:

- using the module materials
- using software that may improve accessibility
- writing mathematics and statistics
- assessment.

The guide does not provide advice on the accessibility of individual modules, or additional or alternative resources that they may provide. This information may be provided in the module description and either on the module website or in the module guide.

General information on the accessibility of the OU’s online resources is available in the accessibility section of The Open University's Computing Guide .

Note that some of the website links in this Accessibility Guide are available only to registered students.

Studying mathematics and statistics

Each mathematics or statistics module involves using a wide range of resources, including some or all of the following:

- module website
- printed units
- scientific calculator
- computer software
- audio, video and tutorial clips (screencasts).

In addition, you will have the opportunity to join in discussions with your tutor and other students, in tutorials and through online forums and email. The tutorials may take place at a study centre or using the University's *Online Rooms* environment, which combine voice, text and images.

By its nature, mathematics is a very visual subject, and most mathematics and statistics modules contain considerable amounts of mathematical notation, graphs and diagrams that you will need to be able to interpret and produce. You may also need to use the computer and mouse to manipulate and enter data. You may need to consider if you are likely to experience problems with activities and assessment questions that involve using your computer, or have high graphical content, particularly if you have difficulties reading print or using a computer for any length of time.

The Open University is committed to making every reasonable effort to make its teaching accessible and to use accessible resources, however some material that is core for a module may not be easily accessible, even if you use assistive technology. You may need a non-medical helper to assist you. If you have any accessibility problems or comments that would help us improve a particular module, please ask your study advisor or tutor to let the module team know. Information about accessibility issues that arise during the life of the module will be posted on the module website.

The module websites

The module websites provide access to a range of resources. These include the Study planner, module resources, any interactive activities and assessment materials. Each website has been checked for accessibility and has an 'Accessibility' link at the top of each page. This provides a link to the OU Accessibility site which has advice on increasing the text size and changing your browser settings, such as colours and fonts. It also includes a link to the accessibility section of the Computing Guide and some useful external links.

The Resources tab may contain a section on Accessibility, which may include figure descriptions of the diagrams and images in the module, and written transcripts of the videos and tutorial clips (screencasts). It may also contain other accessibility resources related to the module. Please refer to the individual module guides for information about this.

Searchable PDF versions of printed material are available in the Resources section of the module website. However, mathematical and statistical content in PDF files is unlikely to be accessible using a screenreader, and older modules may have PDF files that are not searchable. You may need additional help to read these documents.

Modules in mathematics and statistics may require you to spend some or a lot of time online, for example viewing screencasts, completing Interactive Computer Marked Assignments (iCMAs), attending online tutorials or engaging with forums. If you have any accessibility problems relating to this, please ask your study advisor or tutor to let the module team know.

Printed units

The latest information on alternative formats for printed units is available from the Services for Disabled Students website.

Text descriptions are provided for all printed diagrams, charts and graphs that require them. In some cases, you may find it helpful to use the PDF file and enlarge the diagram or graph on your computer screen. An audio graphing calculator can also be used for some activities.

There is free software such as MathTrax that can provide an audio version of some graphs and datasets, and can be used alongside a screenreader. This type of software can be useful if you have difficulty reading text, either printed or on screen.

MathTrax can be downloaded free from the MathTrax homepage.

Notes are available on using MathTrax to plot datasets and graph equations. Further information on MathTrax is also available.

If the effects of your disability make it difficult for you to produce written mathematics, graphs or diagrams, you may like to consider using Efofex, particularly if you are studying Y033 or MU123. This is available free to disabled students under the EmPower programme. Further details are available on the EmPower website or by emailing info@efofex.com for further information.

Scientific calculators

You may need a scientific calculator for your module. Any model that meets the module specification can be used, including computer-based calculators, talking calculators and those with large keys and displays. You can request the loan of a talking scientific calculator from the borrowing small equipment webpage. You can also use the calculator that comes with your computer, if you wish. On the Windows calculator, for example, the menu contains an option to switch the calculator to scientific mode.

You should note that only certain types of calculators are permitted in examinations; further details are available at Calculators permitted in examinations in mathematics and statistics.

Emulators for some types of calculators are available. These enable you to use a calculator on your computer screen and to enlarge the screen and keyboard. An emulator for the Casio fx-82 and Casio fx-85 is available for Windows 7, 8 and 10 from ClassWiz Emulator Subscription v2.00 (90 Day Trial) for Windows. This link includes emulators for many calculators, so please check that you have downloaded and installed a version that meets the exam requirements (for example, the fx-82/85/350EX emulator).

Computer resources

All computer resources have been checked carefully for accessibility, and they can be operated using the keyboard as well as the mouse, unless stated otherwise in the module description. Some components also have their own shortcut keys. For some dynamic computer resources, you may wish to consider the services of a non-medical helper, even if you have assistive technology.

A variety of software applications, some specifically for OU modules and some for more general use, are available for students to download from the Software Downloads section of the Computing Guide.

The Java Access Bridge, which enables Java applications and screenreaders to work together on Windows PCs, is installed as part of the standard Java package; users of assistive technology with Java applications should ensure that the Java Access Bridge is enabled; explicit instructions are given in the Additional software required section of the Computing Guide.

Audio, video and tutorial clips (screencasts)

All audio, video and tutorial clips (screencasts) have transcripts, subtitles and closed captions or alternative versions, unless stated otherwise in the module guide.

Tutorials

Tutorials may be offered either face-to-face or online using the university's conferencing system, *Online Rooms*. Please discuss your requirements with your Student Support Team or with your tutor, particularly if you are deaf or hard of hearing, or use assistive technology such as a screen reader, or if you need particular colour settings.

Writing mathematics using a computer or laptop

This section is aimed at those of you who either need or wish to use a computer or touchscreen to prepare your assignments. Typesetting mathematics is a time-consuming process, but learning how to do it may be a good investment if you are intending to study a lot of mathematics. The Mathematics and Statistics Study website contains information on typing mathematical notation by using either Microsoft Word or the mathematical typesetting language LaTeX. A link to this guidance is under the Skills tab.

Typesetting, even of elementary mathematics, can be tricky. For screenreader users, there are specific issues in that although you may be able to type in maths expressions, the screenreader may not be able to read them back to you. The essential reason why typesetting mathematics is difficult is that written mathematics is two-dimensional in nature. Objects like fractions and matrices have a two-dimensional format and are sometimes stored as images rather than text. This is why screenreaders have difficulty. The good news is that technology is improving all the time and there is software that can help you.

If you use a screenreader and intend to study mathematics beyond Level 1, then you are advised to learn how to use LaTeX, a freely available mathematical typesetting package. This system is long established and it is used by professional mathematicians. The mathematical expressions are typed in a straightforward (plain text) way; experienced blind or/and partially sighted people/students users are likely to read mathematics, as well as write it, in this format. There is further information about Braille and LaTeX on the latex-access sourceforge website.

If you are going to work fully in LaTeX, you might wish to investigate editing software called EdSharp. This text editor has been designed with screenreaders in mind and incorporates facilities for editing LaTeX, HTML and other coding languages.

If you are studying mathematics as part of subject area where Microsoft Word is the normal document software, then you have two main options for entering mathematical content: you may either use Word's inbuilt equation editor, or you may use MathType. If you wish to use Word's inbuilt equation editor, you may like to access this video on Getting started with effective entry of equations in Word. If you obtain version 6.5 or higher of MathType, then you will be able to type in your mathematics using LaTeX notation, but manage the rest of the document in the usual way. You can then toggle between the LaTeX and the maths display by pressing Alt and the backslash key (\) together from within Microsoft Word. This should help you to check your work. There are some further details on the access2science Authoring Math website.

If you decide to use MathType for both your assignments and discussions with your tutor, please check that your tutor can see what you have written. If your tutor does not have MathType installed, then they may not be able to see your mathematical statements. One option is to save your document as a PDF file, so that your tutor can read it easily; this is a reliable method and is recommended.

Some assistive technology, such as ZoomText works best when used with equations produced using Microsoft Word's equation editor. When increasing the font size in the Word document, equations produced in Word format will reflow while those produced with MathType will not.

Some students have found using the commercial software MathTalk helpful for dictating mathematical statements. However, if you decide to use Mathtalk for any assessments, then you must turn off the computer algebra engine and not use the Evaluate, Solve, Integrate or other similar commands. To turn off the engine in MathTalk, please click on Tools, then Computation Setup and then select None in the `Current Engine' drop-down box. The YouTube video titled turning off engine in scientific notebook illustrates how to do this.

Assessment

Assessment in the form of Tutor-Marked Assignments (TMAs) can be submitted by posting a paper copy to your tutor, or electronically submitted using The Open University eTMA system. TMA questions may be in PDF format or accessed directly from the website. In places where graphs or diagrams are used, descriptions will be supplied. You should be able to complete the TMAs successfully with appropriate support from your tutor or helper.

iCMAs are submitted online and may have some interactive content. If you are unable to access the internet (for example, if you are a Student in a Secure environment) or are unable to use the iCMA system for other exceptional reasons, please contact your Student Support Team.

The examinable component for your module may be in the form of an end-of-module assessment (EMA) or an exam. If you have already notified us about a long term health condition, specific learning difficulty or mental health difficulty, you will be contacted by your Student Support Team to discuss any requirements that you may have for the exam.

If you are unable to submit all your assignments or have extra difficulties during the EMA or the exam, you are advised to notify the Exam Board to explain your circumstances. You can do this online by filling in a special circumstances form. If you need further help with the assessment on your module, please contact your tutor.

Summary of web links

The Accessibility section of the Computing Guide has general advice on accessibility for Open University modules, including information on accessibility of equations on module websites. Some external sites also have useful information; for example, Microsoft accessibility has some useful pages on how to access computers and on Windows keyboard shortcuts for accessibility. Similar information is available for Apple computers.

The list below summarises the main websites mentioned in this guide; you can access the university links from your StudentHome page.

We wish you every success with your study of mathematics and statistics. If you do experience any difficulties during your studies, please contact your tutor, who will be able to help you with the assistance of other OU staff in Study Support and Services for Disabled Students.

Useful web links

Description	Web address
Accessibility section of the Computing Guide	https://learn1.open.ac.uk/course/view.php?id=100132&cmid=12929
Accessibility - using a computer for OU study	https://help.open.ac.uk/accessibility-using-a-computer-for-ou-study
Additional software required section of the Computing Guide	https://learn1.open.ac.uk/mod/oucontent/view.php?id=8329&section=5
Borrowing small equipment	https://help.open.ac.uk/borrowing-small-equipment
Calculators permitted in examinations in mathematics and statistics	https://learn2.open.ac.uk/mod/oucontent/view.php?id=1187836

ClassWiz Emulator Subscription v2.00 (90 Day Trial) for Windows	https://edu.casio.com/softwarelicense/index.php#col2
Contact us	http://www.openuniversity.edu/contact-us
Disability Support Form	https://help.open.ac.uk/disability-support-form
Disabled student support	http://www2.open.ac.uk/students/help/topic/disability
EdSharp	https://github.com/EmpowermentZone/EdSharp
Efofex	http://efofex.com/
Getting started with effective entry of equations in Word	http://www.mathcentre.ac.uk/bathmash/Word/index.html
LaTeX − Braille converter	http://latex-access.sourceforge.net
LaTeX in MathType	www.access2science.com/jagqn/WordLatex.html
Mathematics and statistics	https://learn2.open.ac.uk/site/s-maths
MathTalk	http://www.mathtalk.com
MathTrax	https://prime.jsc.nasa.gov/mathtrax/
MathTrax − drawing graphs	http://prime.jsc.nasa.gov/mathtrax/SHOWME/equations.htm
MathTrax − plotting datasets	http://prime.jsc.nasa.gov/mathtrax/SHOWME/data.htm
MathType	www.dessci.com/en/products/mathtype
Module website accessibility	https://help.open.ac.uk/accessibility-using-a-computer-for-ou-study
Overcoming Barriers to study if you have a disability or health condition	http://www2.open.ac.uk/students/help/overcoming-barriers-to-study-if-you-have-a-disability-or-health-condition
Software downloads	https://learn1.open.ac.uk/mod/oucontent/view.php?id=12658
Special circumstances	www.open.ac.uk/assessment/requests/special-circumstances.php
Student support team	https://help.open.ac.uk/your-contacts
Supporting students with disabilities	http://www.open.ac.uk/courses/do-it/disability
Turning off engine in scientific notebook	https://www.youtube.com/watch?v=ch0PU4QXlFc
Typing Mathematical notation	https://learn2.open.ac.uk/course/view.php?id=206217&cmid=1172763
ZoomText	https://www.zoomtext.com/

‹ Accessibility for Mathematics and Statistics modules

Figure descriptions for M303 Book A

Chapter 1

Figure 1.1

This is a diagram of a right-angled triangle. The right angle, at the bottom right, is marked by a little box. Outside the triangle, double-headed arrows parallel to each side indicate the lengths of the three sides. These are labelled x, y and z, with z being the longest side (opposite the right angle).

Figure 2.1

This is a diagram illustrating the first five triangular numbers. Five equilateral triangles formed from small blue dots arranged in rows are shown. In the labels, each T n is written T subscript n. The triangles increase in size from T 1 on the left to T 5 on the right. The number of dots in each triangle equals the triangular number. Each triangle is formed by adding to the previous triangle, as follows.

T 1 consists of one dot.

T 2 has the dot of T 1 and, below it, an additional row of two dots, making three in total. The two new dots are slightly offset from the old dot, one to each side.

T 3 has the two rows of T 2 and, below them, an additional row of three dots, making six in total.

For T 4 a row of four dots is added, making ten dots in the triangle.

For T 5 a row of five dots is added, making fifteen.

In each case the outer dots are joined by lines, to outline the triangle shape.

Figure 2.2

This is a diagram showing five squares, each made up of two right-angled triangles. The shapes are formed from small blue and white dots, aligned in rows and columns. The squares increase in size from 1 by 1 on the left to 5 by 5 on the right. In the labels, each T n is written T subscript n.

The first square has one blue dot, and is labelled T 1 = 1 squared.

The second square has one white dot representing T 1 in the bottom left corner, and three blue dots representing T 2 to complete the 2 by 2 square. Thus the three dots of T 2 are shown as a row of two dots at the top, and one dot at the bottom right (next to the white dot of T 1). The label is T 1 + T 2 = 2 squared.

The third square has the three dots of T 2 in white at the bottom left, and the six dots of T 3 in blue to complete the 3 by 3 square. The label is T 2 + T 3 = 3 squared.

The pattern continues, with T 3 + T 4 = 4 squared, and T 4 + T 5 = 5 squared.

Thus each triangular number forms a right-angled isosceles triangle, with its dots arranged in rows. In each square, the smaller triangular number fills the bottom left, and the larger one fills the top right. The diagonal of the square coincides with the hypotenuse of the larger triangle.

Figure 2.3

This is a diagram showing a square of 81 dots, split into eight triangular octants, alternately blue and white, and a central pale blue dot. Each octant represents T 4. The ten dots of each T 4 are arranged in a right-angled isosceles triangle with four dots along each side.

Each pair of octants fits together to form a 4 by 5 rectangle, half blue and half white. The resulting four rectangles are fitted together to form a big square. Each rectangle occupies one corner of the big square, and they lie alternately longways and crossways, so that each side of the square has length 4 + 5 = 9. This arrangement leaves a hole in the middle, which is filled by one pale blue dot.

Figure 2.4

This is a diagram showing five squares, each formed from rows of blue dots, which are joined up in L shapes. The squares increase in size from 1 by 1 on the left to 5 by 5 on the right. Each is formed by adding an L to the previous square.

The first square has one dot and is labelled 1.

In the second square, the old dot is shown at the top left and three new dots are added below and to the right of the old one, completing a 2 by 2 square. The three new dots form a backwards L shape and are connected by lines. The label is 1 + 3.

The third square consists of this 2 by 2 square plus another, larger, backwards L shape, to the right and below, which adds another five dots. Each L is joined up by lines connecting the dots. The label on the third square is 1 + 3 + 5.

The pattern continues in the other two squares, which are labelled 1 + 3 + 5 + 7 and 1 + 3 + 5 + 7 + 9.

Figure 2.5

This is a diagram showing four regular pentagons, formed from arrangements of dots, representing the second, third, fourth and fifth pentagonal numbers. Each is formed by adding to the previous pentagon.

The first pentagon, on the left, has five dots, one at each corner, joined by straight lines. It is labelled 1 + 4.

The second pentagon has the first pentagon, smaller, at the top, hanging inside a larger pentagon with sides twice as long. The apexes of the small and large pentagons coincide. The upper sides of the small pentagon are extended downwards to be of length 2. These

form the upper sides of the large pentagon, each comprising dot line dot line dot. The other three sides of the large pentagon are added, with a dot halfway along each side as well as at each corner. Thus seven new dots are added to the first pentagon. The label is 1 + 4 + 7.

The third pentagon is formed by repeating this construction. The two nested pentagons hang inside another pentagon with side length three. Its sides comprise dot line dot line dot line dot. Ten dots are added this time, and the label is 1 + 4 + 7 + 10.

The fourth pentagon is formed in the same way, nesting the three previous pentagons inside another pentagon with side length 4. The label is 1 + 4 + 7 + 10 + 13.

Thus at each iteration three new sides are added, whose length is one more than those in the previous iteration. Thus three more dots are added than in the previous step.

Figure 2.6

This is a diagram showing three regular hexagons, formed from arrangements of dots, representing the second, third and fourth hexagonal numbers. The construction is similar to the one for pentagonal numbers.

The first hexagon, on the left, has six dots, one at each corner, joined by straight lines. It is labelled 1 + 5.

The second hexagon has the first one hanging inside it at the top left, so that the top left corners coincide and the top and top left sides of the smaller one extend halfway along the top and top left sides of the larger one. The remaining four sides of the second hexagon, of length 2, are added, each comprising dot line dot line dot. Thus nine new dots are added to the first hexagon. The label is 1 + 5 + 9.

The third hexagon is formed by repeating this construction to obtain three nested hexagons, the largest of side length 3. The label is 1 + 5 + 9 + 13.

Thus at each iteration four new sides are added, whose length is one more than those in the previous iteration. Thus four more dots are added than in the previous step.

Figure 2.7

This is a 3D diagram showing four regular triangular pyramids, formed from layers of balls. Each is formed by adding to the previous pyramid.

The first pyramid has one ball, labelled 1.

The second pyramid has the first ball at the top, and a triangle of three balls below. A line joins each pair of balls, forming a pyramid or tetrahedron. The label is 1 + 3.

The third pyramid has the previous pyramid at the top, and a triangle of six balls, the next triangular number, below. A network of lines marks out the six edges of the pyramid and connections within each layer. The label is 1 + 3 + 6.

The fourth pyramid is formed by adding a fourth triangular layer of ten balls, the fourth triangular number, to make a pyramid with four balls along each edge. The label is 1 + 3 + 6 + 10.

Figure 2.8

This is a 3D diagram similar to Figure 2.7, showing four pyramids, but here the bases are square, not triangular. Each is formed by adding to the previous pyramid.

The first pyramid has one ball, labelled 1.

The second pyramid has the first ball at the top, and a square of four balls below. Lines mark out the eight edges of the pyramid. The label is 1 + 4.

The third pyramid has a square layer of nine balls added below. The label is 1 + 4 + 9.

The fourth pyramid has a square layer of sixteen balls added below. The label is 1 + 4 + 9 + 16.

Figure 3.1

This is a diagram showing fourteen dominoes successively pushing one another over. The dominoes are standing up on their short sides and are drawn as a row of thin rectangles with small gaps between, so that each hits the next if it falls. The five on the right are still upright. The first one on the left is considerably tilted to the right, and is partly resting on the second one, which is a little less tilted, and partly resting on the third one, and so on. The impression is that soon the whole row will fall down. To the right of the dominoes, three dots indicate that the row of dominoes could extend for ever. Underneath the diagram on the left it reads, The first domino is knocked over, and on the right it reads, If the k th domino is knocked over, so too is the k + 1 th.

Figure 3.2

This is a diagram showing eighteen dominoes, illustrating the example of generalised mathematical induction with n 0 = 10. The dominoes are standing up on their short sides and are drawn as a row of thin rectangles, as in Figure 3.1, but this time the gaps between them are not all the same.

On the right of the row are dominoes ten to eighteen, spaced with small gaps. Domino ten is tilted to the right, knocking over domino eleven, which is knocking domino twelve, and so on. The last two are still upright but the impression is that they will fall soon, and three dots indicate that the row could extend to the right for ever and all would fall. Domino ten is labelled P 10 true.

To the left of domino ten, dominoes four to nine are upright, with small gaps. A double-headed arrow below dominoes four to eighteen and beyond is labelled, If P k true then P k + 1 true.

To the left of domino four, dominoes two and three are upright, with large gaps, so neither would knock its neighbour if it fell. Above the upright dominoes two to nine is the label P 2 to P 9 all false.

To the left of domino two, domino one lies flat, labelled P 1 true. Domino one's fall has not affected domino two.

Each P k is written with parentheses, P(k).

Figure 4.1

This is a diagram of a number line, marked in equal intervals of length b. A solid line runs horizontally, representing a section of the real line. At regular intervals from left to right the following points are marked: minus 3 b, minus 2 b, minus b, 0, b, 2 b, 3 b. Then there is an unlabelled point followed by a gap where the line becomes dotted. Then the solid line restarts with points labelled q b, (q + 1) b. The point A is shown between q b and (q + 1) b.

Chapter 2

Figure 1.1

This is a table of the numbers 2 to 100, arranged in order in five rows that end with 20, 40, 60, 80, 100. There is a gap at the top left where 1 is omitted. Most of the numbers are crossed out, using four different types of stroke, as described in the text. The first four primes 2, 3, 5 and 7 are ringed, all near the beginning of the top row. The remaining uncrossed numbers, the primes, are as follows: in the top row 11, 13, 17, 19; in the second row 23, 29, 31, 37; in the third row 41, 43, 47, 53, 59; in the fourth row 61, 67, 71, 73, 79; in the last row 83, 89, 97.

Figure 1.2

This is a table of the numbers 101 to 140, arranged in order in four rows that end with 110, 120, 130, 140. Most of the numbers are crossed out, using four different types of stroke, as described in the text. No numbers are ringed. The uncrossed numbers, the primes, are as follows: in the top row 101, 103, 107, 109; in the second row 113; in the third row 127; in the last row 131, 137, 139.

Figure 1.3

This is a tree diagram with the root, 660, at the top, and three branches descending from it. Each branch has three nodes after the root, joined by downward arrows. Each node consists of a multiplication sum, as follows.

On the first branch: 11 times 60, then 11 times 4 times 15, then 11 times 2 times 2 times 3 times 5.

On the second branch: 15 times 44, then 3 times 5 times 2 times 22, then 3 times 5 times 2 times 2 times 11.

On the third branch: 20 times 33, then 2 times 10 times 3 times 11, then 2 times 2 times 5 times 3 times 11.

Figure 3.1

This is a spiral diagram of the numbers 41 to 53. An arrow points from each number to its successor. All turns are right angles. The path starts at 41 in the middle, then goes up to 42, left to 43, down to 44 and continues down to 45. 45 is in the bottom row, and the spiral continues to the right across the page, so that 46 is below the starting point. It continues right to 47, then up to 48, 49 and 50, which is in the top row. The path then goes left to 51, 52 and 53, where it ends. Thus the diagonal from top left to bottom right consists of 53, 43, 41 and 47.

Figure 3.2

This is a spiral diagram that continues the spiral in Figure 3.1, constructed by placing numbers from 41 to 383 on an anticlockwise path. The composite numbers have been removed, without changing the position of the primes. There are no arrows between numbers, but between successive numbers in the path there are lines of variable length, from one prime to the next. Towards the middle, on the diagonal from top left to bottom right, 53, 43, 41 and 47 appear as before. The spiral now has nineteen rows, and the diagonal consists of the nineteen numbers 347, 281, 223, 173, 131, 97 and 71, then 53, 43, 41 and 47, then 61, 83, 113, 151, 197, 251, 313 and lastly 383. These nineteen numbers are in bold type. After the last number, an arrow points up to indicate that the spiral could be extended.

Figure 5.1

This is a tree diagram with the root, the first pair of rabbits at the beginning, at the bottom left. Across the bottom is an axis marked in months from 0 to 6. Each column represents a month. Each row represents a particular pair of rabbits, represented by its number, in order of birth. The numbers are ringed. Thus the first pair is number one, and appears at every month in the bottom row, to indicate that pair 1 continues to exist each month. Dotted lines run across the rows from number to number.

A birth is represented by an arrow from parent pair to child pair. The first birth, of pair two by pair one, is shown by an arrow from the one in month one to a two in month two. Thus in the column for month two, there are two ringed numbers, number one in the bottom row and number two in a row higher up. The two then appears again in each successive month in its row. Each arrow points up and right, to indicate that the birth occurs at the end of one month and the new rabbits are first counted in the next month.

The other numbers, and the first months when they appear, are as follows.

Three from parents one in month three.

Four from parents one in month four. Five from parents two in month four.

Six from parents one in month five. Seven from parents three in month five. Eight from parents two in month five.

In the last column, month six, numbers nine to thirteen are born from parents 1, 4, 3, 2, 5 respectively.

The complete list of ringed numbers in each column, indicating which pairs exist in each month, is as follows: month zero, number one; month one, number one; month two, numbers one and two; month three, numbers one to three; month four, numbers one to five; month five, numbers one to eight; lastly, month six, numbers one to thirteen.

The diagram has a key, as follows. An arrow to a ringed n denotes the birth of the n th pair of rabbits. A dotted line to a ringed n denotes the continued existence of pair n.

Chapter 3

Figure 2.1

This is a diagram showing a long multiplication sum and how to check it.

On the left is the sum. In the top row is 1 4 8 9, and below it is 7 5 3, aligned so that the units, tens and so on are in columns. A times sign indicates that 1 4 8 9 is to be multiplied by 7 5 3. A horizontal line is drawn under the multiplicands, then the next three rows show 1 4 8 9 being multiplied by 700, 50 and 3 respectively to give 1 0 4 2 3 0 0, then 7 4 4 5 0, then 4 4 6 7. Finally another line is drawn and the three previous rows are added to give the answer 1 1 2 1 2 1 7.

On the right is the check, by summing the digits of each of the multiplicands and calculating the remainder modulo 9. The top row reads 1 + 4 + 8 + 9 = 22 is congruent to 4 mod 9. The second row reads 7 + 5 + 3 = 15 is congruent to 6 mod 9. These totals, 4 and 6, are then multiplied together to give 4 times 6 = 24 is congruent to 6 mod 9. Thus the answer should be congruent to 6 mod 9. Finally, at the bottom next to the answer, the digits of the answer are summed. It reads, Check: 1 + 1 + 2 + 1 + 2 + 1 + 7 = 15 is congruent to 6 mod 9.

Is congruent to is written like equals but with three parallel lines instead of two.

Figure 4.1

This is the annotated solution to the three simultaneous congruences.

The top row lists solutions to the first congruence as follows: x is congruent to 6 mod 13 implies x = 6, 19, 32, 45, 58, 71, 84, 97, dot dot dot. It is annotated with a cloud bearing the following text: increase by 13 until a solution of the second congruence is found.

An arrow runs from the last number in the first row, 97, to the number 97 where it appears in the second row. This is a solution to both the first and the second congruences.

The second row lists solutions to the second congruence as follows: x is congruent to 9 mod 13 implies x = 97, 240, 383, 526, 669, dot dot dot. It is annotated with a cloud bearing the following text: increase by 11 times 13 until a solution of the third congruence is found.

An arrow runs from the last number in the row, 669, to the number 669 where it appears in the third row. This is a solution to the first, second and third congruences simultaneously.

The third row reads as follows: x is congruent to 4 mod 7 implies x = 669, dot dot dot. It is annotated with a cloud bearing the following text: plus any multiple of 7 times 11 times 13.

Implies is written as an elongated equals sign with an arrow head pointing right. Is congruent to is written like equals but with three parallel lines instead of two.

Chapter 4

Figure 1.1

This figure shows three tables, labelled mod 3, mod 5 and mod 7. In each case the top row, labelled A, lists the non-zero residue classes. The other rows list powers of the top row.

The table mod 3 has two rows and two columns. The first row is labelled A, with entries 1, 2. The second row is labelled A squared, with entries 1, 1.

The table mod 5 has four rows and four columns. The first row is labelled A, with entries 1, 2, 3, 4. The second row is labelled A squared, with entries 1, 4, 4, 1. The third row is labelled A cubed, with entries 1, 3, 2, 4. The fourth and last row is labelled A to the fourth, and is a row of ones.

The table mod 7 has six rows and six columns. The first row is labelled A, with entries 1, 2, 3, 4, 5, 6. The second row is labelled A squared, with entries 1, 4, 2, 2, 4, 1. The third row is labelled A cubed, with entries 1, 1, 6, 1, 6, 6. The fourth row is labelled A to the fourth, with entries 1, 2, 4, 4, 2, 1. The fifth row is labelled A to the fifth, with entries 1, 4, 5, 2, 3, 6. The sixth and last row is labelled A to the sixth, and is a row of ones.

Figure 2.1

This figure shows a long division sum on the left, in a long thin column, and on the right, a column of corresponding calculations in terms of the division algorithm.

In the long division of 1 by 7, the dividend 1 point 0 is written near the top, with 7 to its left. A line separates the 7 from the 1 and continues across above the 1. The answer 0.1428571 dot dot dot is written on this line. Digits in the calculation below are aligned according to their decimal place, so that the column veers to the right as it descends.

The first step is divide 1 by 7, giving integer part 0, so record 0 point in the answer line. Next convert 1 to 10 and divide 10 by 7, giving integer part 1. Record 1 in the answer line and subtract 7 from 10 to obtain the remainder 3. Bring down a zero, converting 3 to 30, and divide 30 by 7. And so on. Line by line under 1 point 0 are 7, underline, 30, 28, underline, 20, 14, underline, 60, 56, underline, 40, 35, underline, 50, 49, underline, 10, 7, underline.

The box referred to in the text surrounds 20, 14 and the 6 part of 60 with a dotted line. An arrow points up from the box to the digit 2 in the answer, indicating that 20 divided by 7 gives 2. The remainder 6 is written below and converted to 60.

In the division algorithm version, eight equations are aligned in a column. The top line reads 1 = 0 times 7 + 1. The pattern is quotient times 7 + remainder. The quotients align in a column, forming the digits of the answer. A curving arrow runs from each remainder to the beginning of the next line, indicating multiplication by 10. Thus the remainder 1 becomes 10 in the second line. The second line reads 10 = 1 times 7 + 3. The other six lines read 30 = 4 times 7 + 2, 20 = 2 times 7 + 6 (which corresponds to the boxed calculation on the left), 60 = 8 times 7 + 4, 40 = 5 times 7 + 5, 50 = 7 times 7 + 1, and lastly, 10 = 1 times 7 + 3, the same as the second line.

To the right of each equation is a congruence modulo 7, indicating that the expressions are congruent to successive powers of 10. Thus the second line reads 10 = 1 times 7 plus 3 is congruent to 10, and each line below that finishes similarly: is congruent to 10 squared, is congruent to 10 cubed and so on up to 10 to the seventh. A dotted arrow runs from the remainder 1 in the seventh equation to the 10 at the beginning of the second line, indicating the cycle.

Figure 2.2

This is a diagram showing the three stages of the faro shuffle.

At first the fifty-two cards are in order, from 1 to 52. Each card is shown as a numbered rectangle. The cards are arranged in a single column, numbered 1 to 3, dot dot dot, 50 to 52.

Next the cards are split into two equal columns, numbered 1 to 26 (shown as 1 to 3, dot dot dot, 25, 26) and 27 to 52 (shown as 27 to 29, dot dot dot, 51, 52).

Finally, the cards are again in one column, but interleaved so the numbering is 27, 1, 28, 2, dot dot dot, 51, 25, 52, 26. Thus the pack alternates between cards from the two half packs of the second stage. This alternating pattern is emphasised by displacing the cards from the second half pack a little to the right. The order within the half packs remains the same at each stage.

Arrows from left to right run between the columns, indicating the order of the stages.

Figure 5.1

This is a diagram showing five straight strings of beads laid out in rows, and, on the right, a circular necklace. Each string comprises five beads, represented by ringed letters joined by lines. The letters in the top string are A, A, B, A, C. In each subsequent string the letters cycle one place to the left, so the second string is A, B, A, C, A; the third string is B, A, C, A, A; the fourth string is A, C, A, A, B; the fifth and last string is C, A, A, B, A.

The necklace on the right has the same five beads, joined in a circle. Starting from A at the top and reading anticlockwise, the letters are the same as in the first string, that is A, A, B, A, C. Starting at each of the other beads and reading anticlockwise produces another of the five strings.

Figure 5.2

This is a diagram showing a straight string of four beads, represented as in Figure 5.1. The letters are B, A, B, A.

Figure 5.3

This is a diagram showing a straight string of five beads, represented as in Figure 5.1. The letters are C, A, B, A, A. Reading from right to left, the letters are A, A, B, A, C.

Figure 5.4

This is a diagram showing a straight string of five beads, represented as in Figure 5.1. The letters are A, B, A, B, A. Reading from right to left gives the same sequence of letters.

Figure 5.5

This is a diagram showing twelve pentagons inside circles, arranged in two rows of five, then a row of two. The circles are all the same size, each with five dots evenly spaced around the circumference. The dots are named for the purposes of this description. One dot is directly above the centre. This is the top dot. Below it, two dots at the same level are the middle dots, and two dots closer together on a level near the bottom are the bottom dots. The pentagons are formed by joining the five dots in a closed cycle of straight lines. The last pentagon is the usual regular pentagon with no crossing edges, but the others all have crossings.

In the top row, the first pentagon has left right symmetry. From the top dot, edges run to the middle dots. From the middle dots, edges cross to the bottom dots. A fifth edge joins the bottom dots, completing a distorted figure of eight shape. The other four pentagons in the top row are like the first except that each dot in turn takes the place of the top dot.

In the second row, the first pentagon has left right symmetry. From the top dot, edges run to the bottom dots, then back up to the middle dots, making a W shape. A fifth edge joins the middle dots, crossing the first two edges. The other four pentagons in the second row are like the first except that each dot in turn takes the place of the top dot.

There are two pentagons in the third and last row. The first is a symmetrical five-pointed star, with a point at each dot. The second is the usual regular convex pentagon with no crossing edges.

Figure descriptions for M303 Book B

Chapter 5

Figure 1.1

This figure shows a rectangle with the two lines of reflective symmetry shown. One joins the centre points of the two shorter sides and is marked q subscript 0. The other joins the centre points of the two longer sides and is marked q subscript pi over 2. The rectangle also has an arrow depicting rotational symmetry, marked r subscript pi. The centre of the rectangle is marked as O.

Figure 1.2

This figure shows a rectangle with vertices labelled as follows: top right A, bottom right B, bottom left C, top left D. The centre is marked as O.

Figure 1.3

This figure shows a rectangle with vertices labelled as follows: top right 1, bottom right 2, bottom left 3, top left 4. The centre is marked as O.

Figure 1.4

This figure shows a regular hexagon (six-sided shape) with centre O.

Figure 1.5

This figure shows five identical arcs from a central point. The arcs are regularly spaced around the centre. An arrow indicates the rotational symmetry and is labelled r subscript 2 pi over 5.

Figure 1.6

This figure shows five identical arcs from a central point. The arcs are regularly spaced around the centre. An arrow indicates the rotational symmetry and is labelled r subscript 2 pi over 5. The end points of the arcs are numbered from 1 to 5 in an anticlockwise direction.

Figure 4.1

This figure illustrates the Correspondence Theorem. On the left-hand side, the trivial subgroup is shown connected to a normal subgroup N. N is shown connected to the group G via two different subgroups, H and K. H is shown to be another normal subgroup of G, while K is non-normal.

On the right-hand side of the figure, the subgroup N/N is shown at the bottom and is connected via two subgroups K/N and H/N to the group G/N. H/N is shown as normal in G/N, while K/N is non-normal.

Connecting lines are drawn across the diagram to show the correspondences N/N to N, H/N to H, K/N to K and G/N to G.

Figure 5.1

This figure shows four identical arcs from a central point. The arcs are regularly spaced around the centre. The end points of the arcs are numbered from 1 to 4 in an anticlockwise direction.

Figure 5.2

This figure contains five smaller figures:

(a) has three identical arcs departing from a central point and spaced regularly around that point. The ends of the arcs are labelled 1, 3, 4 in an anticlockwise direction.

(b) has two perpendicular lines of unequal length crossing at their midpoints. The vertical line has endpoints labelled 1 and 2 (top to bottom) and the horizontal line has endpoints labelled 3 and 4 (left to right).

(c) is a single vertical line with endpoints labelled 1 and 2 (top to bottom).

(d) is a rectangle with vertices labelled 1, 4, 3, 2 in an anticlockwise direction with 1 at the top left.

(e) is a triangle with vertices labelled 1, 3, 4 in an anticlockwise direction with 1 at the top.

Figure 5.3

This figure shows a tetrahedron whose vertices are labelled 1, 2, 3, 4. The tetrahedron rests on one of its triangular faces, and the vertices of this face are numbered from 1 to 3 in an anticlockwise direction. The remaining vertex (at the top of the tetrahedron) is numbered 4.

Solution to Exercise 5.13

This figure illustrates the Correspondence Theorem for S4 with V as the normal subgroup. On the left-hand side, the trivial subgroup is shown connected to V, A4 and S4 by a vertical line. On the right-hand side, the subgroup V/V is shown at the bottom and it is connected via a vertical line to the subgroup A4/V and to the quotient group S4/V. Connecting lines are drawn across the diagram to show the correspondences V/V to V, A4/V to A4 and S4/V to S4.

Chapter 6

Figure 3.1

This figure shows an equilateral triangle with lines joining the midpoints of each pair of sides. This divides the triangle into four congruent equilateral triangles. The outer three are numbered 1, 2 and 4 anticlockwise and the middle triangle is numbered 3. The reflection that maps triangle 2 to triangle 4 is labelled s and the rotation through 2 pi over 3 is labelled r.

Figure 3.2

This figure shows an equilateral triangle with a line of reflection from one vertex to the midpoint of the opposite side. This reflection is labelled as s. An arrow indicating the anticlockwise rotation r, through 2 pi over 3, is also shown.

Chapter 7

Figure 1.1

This figure shows a hexagon with a line of reflection through opposite vertices. The reflection is labelled as s and the centre of the hexagon is labelled O.

Figure 1.2

This figure illustrates the Correspondence Theorem for D6. It contains two diagrams.

On the left-hand side of the first diagram, the trivial subgroup is shown connected to the normal subgroup N. N is shown connected to the group D6 via the normal subgroup H. On the right-hand side of the first diagram, the subgroup N bar is shown at the bottom and is connected via H bar to the group D6/N. Connecting lines are drawn across the diagram to show the correspondences N bar to N, H bar to H, and D6/N to D6.

On the left-hand side of the second diagram, the trivial subgroup is shown connected to the normal subgroup N. N is shown connected to the group D6 via the three non-normal subgroups K, K1 and K2. On the right-hand side of the second diagram, the subgroup N bar is shown at the bottom and is connected via K bar,K1 bar and K2 bar to the group D6/N. Connecting lines are drawn across the diagram to show the correspondences N bar to N, K bar to K,K1 bar to K1, K2 bar to K2 and D6/N to D6.

Figure 1.3

This figure shows two sets (one blue and one orange) of four dots (dancers) arranged as opposite pairs (as for a square dance). The orange pairs are horizontally opposite each other and the blue pairs are vertically opposite each other. The dancers are numbered 1 to 8 anticlockwise with number 1 being at the top of the right-hand orange pair.

Figure 1.4

This figure shows the dancers twice.

The left-hand diagram is the same as Figure 1.3, i.e. there are two sets (one blue and one orange) of four dots (dancers) arranged as opposite pairs (as for a square dance). The orange pairs are horizontally opposite each other and the blue pairs are vertically opposite each other. The dancers are numbered 1 to 8 anticlockwise with number 1 being at the top right of the right-hand orange pair.

In the right-hand diagram, the dancers have rotated anticlockwise by pi over two so that the orange pairs are now vertically opposite each other and the blue pairs are horizontally

opposite each other. The dancers are still numbered 1 to 8 anticlockwise but number 1 is now at the top left of the diagram.

Figure 1.5

This figure shows the dancers twice.

The left-hand diagram is the same as Figure 1.3, i.e. there are two sets (one blue and one orange) of four dots (dancers) arranged as opposite pairs (as for a square dance). The orange pairs are horizontally opposite each other and the blue pairs are vertically opposite each other. The dancers are numbered 1 to 8 anticlockwise with number 1 being at the top right of the right-hand orange pair.

The right-hand diagram looks the same as the left with different numbering. The four dancers in the orange set have rotated one place clockwise within their set of four, while the four dancers in the blue set have rotated one place anticlockwise within their set of four. This means that number 1 is now at the bottom of the right-hand orange pair and the dancers are numbered 1, 4, 7, 2, 5, 8, 3, 6 anticlockwise.

Figure 1.6

This figure shows the dancers twice.

The left-hand diagram is the same as Figure 1.3, i.e. there are two sets (one blue and one orange) of four dots (dancers) arranged as opposite pairs (as for a square dance). The orange pairs are horizontally opposite each other and the blue pairs are vertically opposite each other. The dancers are numbered 1 to 8 anticlockwise with number 1 being at the top right of the right-hand orange pair.

The right-hand diagram is the same as the left but with all dancers having moved round the circle by pi (or, in effect, swapped with their opposite number). Thus the dancers are still numbered 1 to 8 anticlockwise but with number 1 now being at the bottom of the left-hand orange pair.

Figure 1.7

This figure shows 16 dancers twice.

In the left-hand diagram there are two sets (one blue and one orange) of eight dots (dancers) arranged as opposite foursomes. The blue foursomes are opposite each other along the top left to bottom right diagonal and the orange foursomes are opposite each other along the top right to bottom left diagonal. The dancers are numbered 1 to 16 anticlockwise, with number 1 being the lowermost of the top right orange foursome.

In the right-hand diagram, the dancers have rotated so that the blue foursomes are now top right to bottom left and the orange foursomes are top left to bottom right. Number 1 is now the uppermost of the top left orange foursome.

Figure 1.8

This figure shows the same 16 dancers twice.

The left-hand diagram is the same as the left-hand diagram in Figure 1.7, i.e. there are two sets (one blue and one orange) of eight dots (dancers) arranged as opposite foursomes. The blue foursomes are opposite each other along the top left to bottom right diagonal and the orange foursomes are opposite each other along the top right to bottom left diagonal. The dancers are numbered 1 to 16 anticlockwise, with number 1 being the lowermost of the top right orange foursome.

The right-hand diagram looks the same as the left with different numbering. The eight dancers in the orange set have rotated one place clockwise within their set of eight, while the eight dancers in the blue set have rotated one place anticlockwise within their set of eight. This means that number 1 is now the lowermost of the bottom left orange foursome and the dancers are numbered 1, 8, 13, 14, 15, 2, 3, 4, 9, 16, 5, 6, 7, 10, 11, 12 anticlockwise.

Solution to Exercise 1.1

This figure shows a hexagon with the rightmost vertex labelled A and the centre labelled O. It then shows six hexagons as follows:

- The first is labelled s with a line of reflection through A.
- The second is labelled rs and shows a line of reflection through the top right-hand/bottom left-hand sides of the hexagon. This moves A one vertex anticlockwise from its original position.
- The third is labelled r squared s and shows a line of reflection through the top right/bottom left vertices of the hexagon. This moves A two vertices anticlockwise from its original position.
- The fourth is labelled r cubed s and shows a line of reflection through the top/bottom sides of the hexagon. This moves A three vertices anticlockwise from its original position (i.e. to the opposite vertex).
- The fifth is labelled r to the fourth s and shows a line of reflection through the top left/bottom right vertices of the hexagon. This moves A four vertices anticlockwise from its original position.
- The sixth is labelled r to the fifth s and shows a line of reflection through the top left-hand/bottom right-hand sides of the hexagon. This moves A five vertices anticlockwise from its original position (i.e. one vertex clockwise).

Chapter 8

Nothing to describe.

Figure and table descriptions for M303 Book C

Chapter 9

Table 3.1

The table has five rows. From top to bottom these are labelled A, A squared, A cubed, A to the power 4, A to the power 5. The data in each column is as follows:

1, 1, 1, 1, 1.

2, 4, 8, 6, 2.

3, 9, 7, 1, 3.

4, 6, 4, 6, 4.

5, 5, 5, 5, 5.

6, 6, 6, 6, 6.

7, 9, 3, 1, 7.

8, 4, 2, 6, 8.

9, 1, 9, 1, 9.

Table 3.2

The table has six rows. From top to bottom these are labelled A, A squared, A cubed, A to the power 4, A to the power 5, A to the power 6. The data in each column is as follows.

1, 1, 1, 1, 1, 1.

2, 4, 8, 7, 5, 1.

4, 7, 1, 4, 7, 1.

5, 7, 8, 4, 2, 1.

7, 4, 1, 7, 4, 1.

8, 1, 8, 1, 8, 1.

Table 5.1

The table has three rows. From top to bottom these are labelled k, 3 to the power k, order of 3 to the power k. The data in each column is as follows.

1, 3, 6.

2, 9, 3.

3, 13, 2.

4, 11, 3.

5, 5, 6.

6, 1, 1.

Solution to Exercise 1.1

The table has three rows. From top to bottom these are labelled n, tau of n, sigma of n. The data in each column is as follows.

1, 1, 1.

2, 2, 3.

3, 2, 4.

4, 3, 7.

5, 2, 6.

6, 4, 12.

7, 2, 8.

8, 4, 15.

9, 3, 13.

10, 4, 18.

Chapter 10

Table 1.1

This is a table indicating the outcome of multiplying together residues and non-residues. Residue times residue is a residue. Residue times non-residue is a non-residue. Non-residue times residue is a non-residue. Non-residue times non-residue is a non-residue.

Solution to Worked Exercise 4.4

The equation given as the solution to this worked exercise illustrates the calculation of the Legendre symbol minus 134 over 229. By 2.2 part c, this is equal to the product of three symbols: minus 1 over 229, times 2 over 229, times 67 over 229. We consider each symbol in turn.

By 2.2 part e, the symbol minus 1 over 229 is equal to 1. By 3.2, the symbol 2 over 229 is equal to minus 1. By the Law of Quadratic Reciprocity, the symbol 67 over 229 is equal to the symbol 229 over 67. Therefore, the whole equation is equal to minus 1 times the symbol 229 over 67.

By 2.2 part a, this is equal to minus one times the symbol 28 over 67, and by 2.2 part c, this is equal to minus one times the symbol 4 over 67, times the symbol 7 over 67. By 2.2 part b, the symbol 4 over 67 is equal to 1. By the Law of Quadratic Reciprocity, the symbol 7 over 67 is equal to minus 1 times the symbol 67 over 7.

Therefore, the two minus ones cancel, and the whole equation is equal to 67 over 7. By 2.2 part a, this is equal to the symbol 4 over 7, which by 2.2 part b is equal to 1. Therefore, the answer is 1.

Figure 4.1

This is an illustration of the lattice used in the proof of the Law of Quadratic Reciprocity. The figure shows the rectangular region of the plan bounded by the coordinates 0 comma 0, p over 2 comma 0, p over 2 comma q over 2, and 0 comma q over 2. Within this rectangular region, each lattice point A comma b is marked by a dot, where each of A and b is an integer. The straight line y equals x times q over p runs through the rectangle from the bottom left corner at 0 comma 0, to the top right corner at p over 2 comma q over 2. This line divides the rectangle into two parts. The upper part is labelled A, and the lower part is labelled b. The point k comma 0 is marked on the horizontal axis, and the lattice points that lie in region B and whose first coordinate is k have been highlighted. The number of these is marked as equal to the integer part of k times q over p.

Chapter 11

Solution to Worked Exercise 3.2

You may find it useful to ask a sighted helper to assist you with this worked exercise.

Solution to Exercise 3.1

Part a: the result of dividing f of x equals x cubed minus 2 x squared plus 3 x minus 1 by g of x equals x minus 1. Find this by first subtracting x squared times g of x from f of x. This leaves minus x squared plus 3 x minus 1. From this, subtract minus x times g of x, to be left with 2 x minus 1. Finally, subtract 2 times g of x from this to be left with the number 1, which must be the remainder, r of x. The quotient, q of x, is then x squared minus x plus 2.

Part b: the result of dividing f of x equals 2 x to the power 4 minus x plus 1 by g of x equals x squared plus 1. Find this by first subtracting 2 x squared times g of x from f of x. This leaves minus 2 x squared minus x plus 1. From this, subtract minus 2 times g of x. This leaves minus x plus 3, which must be the remainder, r of x. The quotient, q of x, is then 2 x squared minus 2.

Part c: the result of dividing f of x equals 3 x cubed minus 2 x squared plus 1 by g of x equals 2 x plus 1. Find this by first subtracting 3 over 2 x squared times g of x from f of x. This leaves minus 7 over 2 x squared plus 1. From this, subtract minus 7 over 4 x times g of x. This leaves 7 over 4 x plus 1. From this, subtract 7 over 8 times g of x. This leaves one eighth, which is the remainder r of x. The quotient, q of x, is 3 over 2 x squared minus 7 over 4 x plus 7 over 8.

Chapter 12

Figure 1.1

This is an illustration of Pythagoras' Theorem. There is a right-angled triangle with a hypotenuse of length z, and two other sides of lengths x and y. Attached to each side is a square of the same side length, so that the square attached to the hypotenuse has area z squared, and the squares on the other two sides have areas x squared and y squared respectively.

Figure 3.1

This is an illustration showing the nearest Gaussian integer to a complex number z. The figure comprises two graphs of the complex plane.

The graph on the left shows the Gaussian integers marked as a square matrix of dots on the complex plane. The complex number z is also marked, as is the nearest Gaussian integer q to this number z.

The graph on the right zooms in on the portion of the complex plane containing the numbers z and q. Around the point z are four Gaussian integers, forming the corners of a unit square that contains z. The corner that is closest to z is labelled q. Between z and q a line has been drawn, and this is shown as the hypotenuse of a right-angled triangle whose other two sides are parallel with the real and imaginary axes. Each of the two shorter sides of this triangle is shown to have length less than or equal to one half. Thus the length of the hypotenuse, which is equal to the value of the complex modulus of z minus q, is less than or equal to the square root of 2 divided by 2.

Solution to Exercise 3.1

The figure provided as part of the solution to this exercise is an illustration showing the nearest point q in Z [square root minus 2] to a complex number z. The diagram is similar to Figure 3.1. On the left is a zoomed-out version of the complex plane, with the points of Z [square root minus 2] marked as a rectangular matrix of dots. On the right is a graph that zooms in on the region containing z and q. Around the point z are four elements of Z [square root minus 2] that represent the corners of a rectangle. The corner that is closest to z is labelled q. Between z and q a line has been drawn, and this is shown as the hypotenuse of a right-angled triangle whose other two sides are parallel with the real and imaginary axes. The length of the side parallel with the real axis is shown to have length less than or equal to one half. The length of the side parallel with the imaginary axis is shown to have length less than or equal to the square root of 2 divided by 2. Thus the length of the hypotenuse, which is equal to the value of the complex modulus of z minus q, is less than or equal to the square root of 3 divided by 2.

Figure descriptions for M303 Book D

Chapter 13

Figure 1.1

A representation of a function f from one set to another. The two sets are shown as separate shaded ovals, labelled domain A and codomain B. A thick curved arrow from the domain to the codomain is labelled f. A thin straight arrow maps a point x in A to the point f of x, written f(x), in B.

Figure 1.2

A graph of the modulus function. The x and y-axes, and the origin 0, are labelled. The graph consists of two straight lines like a wide letter V, meeting at the origin. For negative x, the graph slopes up to the left from the origin at 45 degrees. For positive x, it slopes up to the right from the origin at 45 degrees. The graph is labelled f of x = mod x. mod x is written as x between two vertical lines.

Figure 1.3

A graph illustrating the inequality mod x less than 2. The x and y-axes, and the origin 0, are labelled. The x-axis runs across the bottom and the y-axis runs up the middle. As in the previous figure, the graph of mod x is shown by a pair of solid straight lines in a V shape, meeting at the origin. The point 2 on the y-axis is marked. Dotted lines run left and right across from 2 to meet the graph, then run down to the x-axis, meeting the x-axis at minus 2 and plus 2. Thus a rectangle is formed, with its base on the x-axis between minus 2 and plus 2. This section of the x-axis gives the solution set of the inequality. The points x = minus 2 and x = plus 2 are shown as hollow dots rather than solid dots, to indicate that they are not in the set. The set consists of the open interval between them.

Figure 1.4

A diagram illustrating the open interval mod x less than c. Part of the real line is shown as a horizontal line – the x-axis – with points minus c, the origin 0, and plus c, equally spaced from left to right. The right-hand end of the real line is capped with an arrow, indicating that the positive direction is to the right. Plus and minus c are marked by hollow dots, to indicate that they are not in the interval, whereas the origin is marked by a solid dot. The line is darkened between minus c and plus c to denote the interval. The interval is labelled mod x strictly less than c.

Figure 1.5

A diagram illustrating the open interval mod x minus b less than c. Part of the real line is shown as a horizontal line – the x-axis – with points b minus c, b, and b plus c equally spaced

from left to right. The right-hand end of the line is capped with an arrow, indicating that the positive direction is to the right. The points b minus c and b plus c are marked by hollow dots, to indicate that they are not in the interval, whereas b is marked by a solid dot, in the centre of the interval. The line is darkened between b minus c and b plus c to denote the interval. The interval is labelled x is an element of (b minus c comma b plus c).

Solution to Exercise 1.5

A diagram showing the solution set of the pair of inequalities. Part of the real line is shown as a horizontal line – the x-axis – with integer points from minus 4 to plus 4 marked along it. The right hand end of the line is capped with an arrow, indicating that the positive direction is to the right. Above the line the solution set to one inequality, mod x + 1 less than 3, is shown by a large horizontal curly bracket stretching from minus 4 to plus 2. Below the line the solution set to the other inequality, mod x minus 1 less than 3, is shown by a bracket from minus 2 to plus 4. The endpoints of the intervals are marked by hollow dots to show they are not in the intervals. The overlap between the two sets, where both inequalities hold, from minus 2 to plus 2, is shown by darkening of that part of the line.

Figure 1.6

A diagram illustrating distance between two points on the real line. Part of the real line is shown as a horizontal line, with points lower-case A and b marked. The right-hand end of the line is capped with an arrow, indicating that the positive direction is to the right. Above the line, parallel to it, is a double-headed arrow from A to b, labelled mod of b minus A. Thus this definition of distance coincides with the everyday meaning.

Figure 1.7

A graph of the first five terms in the sequence one over n. The horizontal axis at the bottom is labelled n, and marked in integers from the origin to 5. The vertical axis is labelled A subscript n, and marked in fractions one fifth, one quarter, one third, a half, and 1. The first five terms of the sequence are marked at points (1, 1) to (5, one fifth). Dotted lines run across and down from each point to the axes. The effect is descending steps, getting progressively shallower, with a big step down from one to a half, and a little step down from one quarter to one fifth.

Figure 1.8

A graph illustrating a sequence that converges to zero, showing how the later terms are confined to a band from minus epsilon to plus epsilon. The horizontal axis, across the middle, is labelled n, and marked in integers from the origin up to capital N + 2. Between 4 and capital N there is a small vertical gap in the band and a wiggle in the n-axis. This corresponds to dot dot dot in a list of values. The vertical axis on the left is labelled A subscript n, with the points minus epsilon, zero, and plus epsilon marked. A shaded horizontal band runs across the graph, from the level minus epsilon up to the level plus epsilon. Terms of the sequence A n, denoted by lower-case A subscript capital N, are

shown. They form an irregular pattern. The point A 1 is above the band, A 2 is within the band below the axis, A 3 is below the band, A 4 is just inside its upper boundary, A capital N is just above it. The two terms with n greater than capital N lie within the band, both below the n-axis.

Figure 1.9

A graph illustrating the sequence minus one to the n, showing that the terms are not confined to a band from minus a half to plus a half. The horizontal axis, across the middle, is labelled n, and is marked in integers from the origin up to 4. The vertical axis on the left is labelled A subscript n, and is marked with the points minus 1, minus epsilon = minus a half, zero, epsilon = plus a half, and plus 1. A shaded horizontal band runs across the graph, from the level minus a half, that is minus epsilon, up to the level plus a half, that is plus epsilon. The first four terms of the sequence A n, denoted by lower-case A subscript n, are shown. The point A 1 = minus 1 lies below the band. The point A 2 = plus 1 lies above the band. This pair of values then repeats, so that the odd values form a horizontal row below the band, and the even values form a horizontal row above the band.

Figure 1.10

A graph illustrating a sequence that converges to a limit l, showing how the later terms are confined to a band from l minus epsilon to l plus epsilon. The horizontal axis, across the bottom, is labelled n, and marked in integers from the origin up to capital N + 2. The vertical axis is labelled A subscript n, and is marked with the points zero, l minus epsilon, l, and l plus epsilon. A shaded horizontal band runs across the graph, from the level l minus epsilon up to the level l plus epsilon. A dotted horizontal line runs across the middle of the band at the level of the limit l. Between 3 and capital N there is a small vertical gap in the band and a wiggle in the n-axis. This corresponds to dot dot dot in a list of values. Terms of the sequence A n are shown, denoted by lower-case A subscript n. They form an irregular pattern. The point A 1 is above the band, A 2 is within the band above the limit line. Then A 3 is below the band, A capital N is just inside the bottom of the band. The two terms with n greater than capital N lie within the band: the first below the limit line and the second above it.

Figure 1.11

This diagram illustrates the argument used to verify that the sequence A n given by A n equals n is divergent. Given a fixed number L, the diagram shows that for any capital N there is a lower-case n larger than capital N for which the term of the sequence A n is strictly larger than the one plus the modulus of l.

The diagram shows the upper right quadrant of a plane with horizontal and vertical axes marked, and has a horizontal shaded strip parallel to the horizontal axis at height the modulus of L and of width one. On the horizontal axis, the point lower-case N is shown, where lower-case N is chosen to be one greater than the maximum of capital N and the least integer greater than or equal to L. The diagram shows that for this value of lower-case N the

term of the sequence A subscript lower-case N lies above the shaded strip, meaning that it is at least distance one away from L.

Figure 1.12

This diagram consists of two graphs. The first graph shows a monotonic increasing sequence that is bounded above by a fixed real number lower-case B subscript U. The sequence is represented by six points that are equally spaced horizontally, but increase in height from left to right. They approach, but do not meet, the horizontal line at B subscript U.

The second graph shows a monotonic decreasing sequence that is bounded below by a fixed real number lower-case B subscript L. The sequence is represented by six points that are equally spaced horizontally, but decrease in height from left to right. They approach, but do not meet, the horizontal line at B subscript L.

Figure 1.13

A graph illustrating the sequence minus 1 to the n, showing that the values split into two subsequences with different limits. The horizontal axis, across the middle, is labelled n, and is marked in integers from the origin up to 3, then, after a wiggle, 2 n minus 1, and 2 n. The wiggle corresponds to dot dot dot in a list of values. The vertical axis is labelled A subscript n, and is marked with the points minus 1, zero, and plus 1. The terms of the sequence lie alternately at the levels minus one and plus one. Each term A n is denoted by lower-case A subscript n. The odd terms, A 1, A 3, and A 2 n minus 1, lie on a dotted horizontal line below the axis, at level minus 1. The even terms, A 2, and A 2 n, lie on a dotted horizontal line above the axis at level plus 1. These two dotted lines represent the two constant subsequences. There are wiggles in both dotted lines, aligned with the wiggle in the n-axis.

Figure 2.1

Two graphs, the first showing the continuous function f 1 of x, the second showing the function f 2 of x, which has a discontinuity at zero. First is a description of the continuous graph. The horizontal axis, across the bottom, is labelled x, with the origin marked in the middle. The vertical axis, up the middle, is labelled y. The graph of the function appears as a continuous symmetrical curve in the shape of a cup, descending from the left towards the origin, passing through the origin, then ascending. The curve is labelled y = f subscript 1 (x). The axes and origin of the second graph are as in the first one. The graph of the function is in two parts, both horizontal. On the left it coincides with the x-axis, up to the origin. The point at the origin is marked with a solid dot to indicate that f 2 of zero = zero. On the right the graph is a horizontal line at the level y = 1, running across from the y-axis. The point x = 0, y = 1 is marked as a hollow dot to indicate that this endpoint is not included in the line. Thus the graph jumps up by one unit at x = 0. The curve is labelled y = f subscript 2 (x).

Figure 2.2

A graph showing the discontinuous function f 2 of x, as in the previous figure, with additionally the sequence 1 over n. On the left, the graph coincides with the x-axis. On the

right, it follows the horizontal line y = 1. The value of f of 0 is 0, indicated by a solid dot at the origin. The graph jumps up by one unit at x = 0. To the right of the origin, on the x-axis, are marked the fractional points one quarter, one third, a half, and 1. Vertical dotted lines from each of these four points run up to meet the graph at y = 1. Thus looking from right to left the fractional points on the x-axis approach the origin. But the corresponding y values are all 1, and do not approach f of 0.

Figure 2.3

A graph showing f o x = cosine of pi over x. It is not immediately clear if the function is continuous at the origin, because the graph oscillates so much that the picture is blurred there. The x-axis, across the middle, extends between about minus 2 and plus 2, with minus 1 and plus 1 marked. The y-axis, up the middle, has minus 1 and plus 1 marked. The graph is a wave, with symmetry about the y-axis, with each oscillation peaking at plus 1 and descending to minus 1. In detail from left to right, the graph starts at the x-axis at x = minus 2, then descends to minus 1 at x = minus 1. It then rises sharply to plus 1 at x = minus a half. The oscillations then become increasingly close together, so that near the y-axis they are too close to distinguish. To the right of the y-axis the pattern continues symmetrically, with the oscillations becoming more widely spaced. Beyond the blurred section there are clear peaks at x = a quarter and x= a half. The graph is labelled y = f of x.

Figure 2.4

A graph showing the same function f of x as in the previous figure. Only the part from x = 0 to 1 is shown, and the horizontal scale is larger so that more oscillations can be clearly distinguished. The first three points of the sequence x n are also shown. The x-axis, across the middle, has 0 and 1 marked. The y-axis, up the left, has minus 1, 0 and plus 1 marked. Near the y-axis the oscillations are too close together to distinguish, but gradually get further apart so that peaks at x = an eighth, a sixth, a quarter and a half can be seen, all reaching y = 1. Troughs reaching y = minus 1 lie in between. The points x 1 = a half, x 2 = a quarter, and x 3 = a sixth are marked, all coinciding with peaks. A dotted line runs across from the y-axis at level 1, connecting the peaks. Dotted lines run from each x n up to this line, to indicate that f of x n = 1 for each n. The graph is labelled y = f o x.

Figure 2.5

A graph showing the function f of x = x times cosine of pi over x, from x = minus 1 to x = 1. The graph is a wave, confined between the diagonal lines y = x and y = minus x, which are also shown, dotted, in the figure. These two lines form an X shape, crossing at the origin. As in the previous figure, the oscillations become increasingly close together near the origin, but the peaks and troughs are no longer at plus and minus 1. Instead, they are on the diagonal lines. In detail, from left to right, at x = minus 1 the graph touches the upper branch of the X shape, then descends to touch the lower branch at minus a half, minus a half, then rises steeply to touch the upper branch at minus a third, a third, and so on. Near the y-axis the oscillations are too close together to distinguish, but it is clear that all points are very close to the origin here, as the branches of the X shape converge towards their crossing. On

the right of the y-axis the oscillations widen again. Peaks on the upper branch include points a quarter, a quarter, and a half, a half. Peaks on the lower branch include points a third, minus a third, and 1, minus 1. The graph is labelled y = f of x.

Figure 2.6

A graph illustrating the restriction rule, showing the function y = x squared passing through a shaded vertical strip. The x-axis, across the bottom, is marked in integers, with 0, 1, 2, and 3 labelled. The y-axis runs up the middle with 5 and 10 labelled. The vertical strip extends from x = 0 to x = 3. The curve y = x squared is cup shaped, descending from the left to the origin then ascending symmetrically through the strip and beyond. The curve leaves the strip at the level y = 9. The part of the curve inside the strip is a solid line to indicate the restricted function, whereas outside the strip it is a dotted line. The curve is labelled f of x = x squared.

Figure 2.7

A diagram illustrating the composition of two functions f and g. Two paths take different routes from set capital A on the left to set capital D on the right. The upper path comprises a thick arrow labelled f, which starts at A and runs up and right to its image set, labelled f of capital A is a subset of capital B. From here, a second thick arrow, labelled g, runs down and right to D. The lower path comprises a single thick arrow. From letter A this thick arrow, labelled g composed with f, runs to D by curving down and up again. Composed with is written as a little circle.

Figure 2.8

A graph illustrating the Intermediate Value Theorem for k = 0. A function y = f of x is shown with f of A on the left below the x-axis, and f of b on the right above the x-axis. The graph has several ups and downs but we focus on the existence of the point x = c where it crosses the x-axis. In more detail, the x-axis runs across the middle, with points marked from left to right: zero, A, c and b. The y-axis runs up the left, with f of A near the bottom, the origin in the middle, and f of b near the top. The curved continuous graph of f of x is drawn for values of x from A to b. Dotted lines run from the endpoints to the axes, emphasising that f of A is below and f of b is above. Following the curve from the point A, f of A, it has several ups and downs, going lower than f of A and higher than f of b, and crossing the x-axis three times, before reaching the point b, f of b. The second point where it crosses the x-axis is at x = c. The curve is labelled y = f of x. In its middle section the curve is represented by a dotted line, to indicate that we may not know values here but we do know that c exists. Elsewhere the line is solid.

Figure 2.9

A graph of the function y = 1 over x. It has two separate continuous curved branches which both approach the y-axis when x is small, but never reach it. The origin is shown in the centre, with the x and y-axes forming a cross. The branches are symmetrical, and are confined to the lower left and upper right quadrants. On the x-axis the points minus 1 and

plus 1 are marked near the edges. The scale on the y-axis is not marked but it is smaller. Following the graph from left to right, the first branch starts just below the x-axis, then curves gradually away and down, becoming steeper, towards the lower y-axis, ending heading almost straight downwards just to the left of the y-axis. The second branch starts at the top just to the right of the y-axis and runs steeply down, then curving away from the y-axis becoming less steep, and ending on the right, almost horizontal, just above the x-axis. The curve is labelled y = f of x.

Figure 2.10

A graph labelled y = f of x. The general impression is of a continuous curve, which has small irregular fluctuations for values of x below about 2, then increases steadily for higher values. The x-axis is marked at integer intervals from minus 3 to 3. The y-axis runs up the middle, with 10 and 20 marked. Following the graph from left to right, the curve starts a little above the x-axis at around x = minus 3, then dips a little below, so that y is negative throughout the interval from x = minus 2 to minus 1. The curve crosses the x-axis into positive values again a little to the right of minus 1, then peaks near the y-axis and dips again slightly before settling to a steady ascent.

Figure 2.11

The same as the previous graph, but this time the key features are the lowest and highest points. The lowest point is in the dip that includes the interval from minus 2 to minus 1. It appears to be about halfway along this interval. The highest point is on the extreme right, beyond x = 3, where the graph is ascending. Here it rises higher than the small peak that occurs near the y-axis.

Figure 3.1

Two 3D diagrams, showing the surfaces f and g. Shading and perspective are used on the page to give an impression of three dimensions, but this description is of the 3D originals. Each has three mutually orthogonal axes labelled as follows: x 1, a horizontal axis pointing towards us; x 2, a horizontal axis pointing right; and x 3, an axis pointing vertically up. Each x n is written x subscript n. The origin 0 lies in the middle. Both surfaces appear to be continuous. Firstly, the surface f is described. The impression is of the symmetrical upper section of an eggshell, with axis along the vertical axis. The lower boundary consists of four rounded arches, with peaks over the four horizontal axes, coming down to four sharp points, one in each quadrant below the horizontal plane. The apex of the surface is above the plane. The point (3, 2) is labelled, lying in the right front quadrant of the horizontal plane. Directly above it on the surface lies f of (3, 2), which equals 7. The surface is labelled f of (x 1, x 2). Secondly the surface g is described. The impression is of a waving piece of cloth, approximately square, which passes through the origin. In the horizontal plane it lies along the two axes, but in the four quadrants the four quarters of the surface wave up or down. From a low point in the left back quadrant, the perimeter rises in an arch towards the front, peaking at the x 2 axis and descending in the left front quadrant. The second side rises from left to right, to a high point in the right front quadrant. The third side dips to the x 2 axis and

rises to a peak in the right back quadrant. The square is then completed by a descent from right to left. The surface is labelled g of (x 1, x 2).

Figure 3.2

Two diagrams, each showing a sequence in the plane. The sequences A n and b n are written lower-case A subscript n and b subscript n. Firstly the sequence given by A n = cosine n, sine n is described. The axes are shown, with the origin in the centre, and a circle with radius 1 centred on the origin. The points minus 1 and plus 1 on each axis are marked, where the circle cuts the axes. These are not part of the sequence. The first seven points of the sequence are marked as points on the circle. Moving anticlockwise from the positive horizontal axis, the point A 1 is about one sixth of the way round, that is one radian, in the upper right quadrant. The other points then appear at one radian intervals, so that A 2 is a little beyond the upper vertical axis, then A 3 is nearly at x 1 = minus 1, then A 4 is below the horizontal axis, and the next two are in the lower right quadrant, with A 6 nearly at 1 on the horizontal axis. The seventh point has coordinates cosine 7, sine 7, and 7 radians is more than a complete rotation, so the sequence starts a second turn around the circle, with A 7 a little before A 1. Secondly the sequence b n = 1 over n, n is described. The horizontal x-axis across the bottom is marked at fractional intervals from the origin at a third, a half, and 1. The vertical axis on the left is marked at integer intervals 1, 2, 3. The first three terms of the sequence are shown, lying on the curve f of x = 1 over x. From right to left, b 1 is the lowest at (1, 1). The curve then ascends, with increasing slope, through b 2 to b 3 at (a third, 3). The curve extends smoothly a little way beyond the end points, giving the impression that the sequence continues to ascend and slowly approaches the vertical axis as n increases.

Figure 3.3

A graph illustrating how to calculate the Euclidean distance between two particular points in the plane, using a right-angled triangle. The horizontal axis, labelled x subscript 1, is across the middle. The vertical axis, labelled x subscript 2, is up the left side. Points (3, minus 1), and (7, 2) are marked and joined with a solid diagonal line. Dotted lines run right from the lower point and down from the upper point, forming a right-angled triangle with the solid line as the hypotenuse. Double-headed arrows parallel to each side of the triangle mark their three lengths, and are labelled with the three calculations, as follows. Length of the base is 7 minus 3 = 4. Height is 2 minus minus 1 = 3. Hypotenuse is the square root of 4 squared plus 3 squared = 5. Thus the Euclidean distance between the two points is 5, the length of the hypotenuse.

Solution to Exercise 3.1

A graph illustrating the Euclidean distance between any two points in the plane, using a right-angled triangle. The horizontal axis, labelled x subscript 1, is across the bottom. The vertical axis, labelled x subscript 2, is up the left side. Points bold A with coordinates A 1 and A 2,and bold b = b 1, b 2, are marked and joined with a solid sloping line. The coordinates are written using subscripts, for example A 1 as lower-case A subscript 1. Point b is higher and to the right of point A. Dotted lines run right from the lower point and down from the upper

point, forming a right-angled triangle with the solid line as the hypotenuse. The three sides of the triangle are labelled with their three lengths, as follows. Length of the base is b 1 minus A 1. Height is b 2 minus A 2. The length of the hypotenuse is the Euclidean distance d 2 of A, b, written d superscript (2) (bold A, bold b).

Figure 3.4

A graph illustrating that when a line is translated, the distance between its end points is unchanged. Horizontal and vertical axes x subscript 1 and x subscript 2 are shown, crossing at the origin. Points bold A and bold b are marked near the top right, with a higher and to the right of b. The points are joined by a solid sloping line. Two copies of this line, with the same length and gradient, are shown. One runs up and right from the origin, with its right end labelled A minus b. The other runs down and left from the origin, with its left end labelled b minus A. Thus the two copies form one straight line. The distance from A to b is the same as the distance from A minus b to the origin, and the same as the distance from b minus A to the origin.

Figure 3.5

A diagram illustrating a sequence of points in the plane converging to a limit l. The points of the sequence A n are written as lower-case bold A subscript n. A circle with centre l, radius epsilon is shaded. The sequence begins outside the circle, with an irregular pattern of points A 1, then A 2, then A 3, followed by a dotted path to A capital N, which lies just outside the circle. Further terms of the sequence all lie inside the circle. Points A capital N + 1 and A capital N + 2 are marked, getting nearer to the centre. Then another dotted path runs from A capital N + 2 to the centre l. The radius is indicated by a double-headed arrow from the centre to the circumference, labelled epsilon.

Figure 3.6

A 3D diagram showing the surface f, which looks similar to the eggshell surface shown in Figure 3.1, but the other way up, that is the bottom of an eggshell. Shading and perspective are used on the page to give an impression of three dimensions, but this description is of the 3D original. Three mutually orthogonal axes are labelled as follows: x 1, a horizontal axis pointing towards us; x 2, a horizontal axis pointing right; and x 3, an axis pointing vertically up. Each x n is written x subscript n. The origin 0 lies in the middle. The surface appears to be continuous. The impression is of the symmetrical bottom of an eggshell, with tip at the origin and axis along the vertical axis. The upper boundary consists of four rounded dips, with lowest points over the four horizontal axes, rising up to four sharp peaks, one in each quadrant above the horizontal plane. The surface is labelled f of (x 1, x 2) = x 1 squared plus x 2 squared.

Figure 3.7

A graph illustrating the projections of a point in the plane onto the two coordinate axes. The horizontal axis labelled x 1 runs across the bottom. The vertical axis labelled x 2 runs up the

left. The point bold A is near the top right, with coordinates A 1 and A 2. Dotted lines run down and left from A, meeting the x 1 axis at A 1 and the x 2 axis at A 2. Thick curved arrows from point A represent the two projection functions and are labelled p 1 and p 2. p 1 maps A to A 1 and p 2 maps A to A 2. Coordinates are written with subscripts, for example A 1 as lower-case A subscript 1.

Figure 3.8

A 3D diagram showing the surface f, which resembles a square cloth draped in a ridge, with caves in its sides. Shading and perspective are used on the page to give an impression of three dimensions, but this description is of the 3D original. Three mutually orthogonal axes are labelled as follows: x 1, a horizontal axis pointing towards us; x 2, a horizontal axis pointing right; and x 3, an axis pointing vertically up. Each x n is written x subscript n. The origin 0 lies in the middle. The top of the ridge runs horizontally above the x 1 axis. The surface slopes down on each side to reach below the horizontal plane. The slopes are curved, with cave shapes hollowed out on each side of the ridge, so that most of the vertical axis is uncovered. It is difficult to see exactly what happens at the origin, in the depths of the cave.

Solution to Exercise 3.4

[Missing figure by Exercise 3.4]

A 3D diagram showing the graph of the function g given in the exercise. It consists of half of the z = 0 plane together with half of the z = 1 plane, with the step between the planes occurring along the line x 1 = x 2. The line x 1 = x 2 is included in the half plane that lies at height zero.

Figure 3.9

Same as Figure 3.6.

Figure 3.10

A graph illustrating that the surface g of x is a continuous plane that passes through the origin. Three mutually orthogonal axes are labelled as follows: x 1, a horizontal axis pointing towards us; x 2, a horizontal axis pointing right; and x 3, an axis pointing vertically up. Each x n is written x subscript n. The origin lies in the middle. The front part of the plane g, nearest to us, is above the horizontal plane x 3 = 0. The plane then slopes gently down away from us, passing through the origin, and the back part of the plane g is below the horizontal plane. The surface is labelled g of (x 1, x 2) = 2 x 1 minus 3 x 2.

Figure 3.11

A 3D graph of the surface h of x, which looks rather similar to the eggshell surface f of x described previously, but is less rounded, looking more like a coffee filter paper than an eggshell, and has a sharp point at the bottom at the origin. Three mutually orthogonal axes

are labelled as follows: x 1, a horizontal axis pointing towards us; x 2, a horizontal axis pointing right; and x 3, an axis pointing vertically up. Each x n is written x subscript n. The origin 0 lies in the middle. The surface is wholly above the horizontal plane, rising up from the origin to form a symmetrical cup with axis along the vertical axis. The upper boundary consists of four rounded dips, with lowest points over the four horizontal axes, rising up to four sharp peaks, one in each quadrant above the horizontal plane. The surface is labelled h of (x 1, x 2) = square root of x 1 squared plus x 2 squared.

Figure 3.12

A diagram of a triangle, illustrating the Triangle Inequality. The vertices are labelled lower-case A, b and c. The sides are labelled with their lengths, the Euclidean distances between the end points. Side A b is labelled d superscript (2) (bold A, bold b) and the other two sides similarly. The key point to notice is that the distance from A to c via b is longer than the distance from A to c directly along A c.

Chapter 14

Figure 1.1

This figure illustrates that a sequence of points in the plane that converges to the point (one half, 0) strictly from the left gets mapped by f to A sequence of points that does not converge to f of (one half ,0) = (1, 0). It shows two copies of the upper right quadrant of the plane, one above the other. The top copy represents the domain of f and the bottom the codomain of f. In the top copy, a sequence labelled x k that converges along the x 1 axis to the point (a half, 0) is shown. In the bottom copy of the plane, the sequence that results after applying f to this sequence is indicated: the image of any point x k under f is just the origin, (0, 0), and so the image sequence does not converge to the image of the point (one half, 0): f (one half, 0) is (1, 0).

Figure 1.2

The graph of the function g in the case when n is equal to two. It is a 3D graph with axes x 1 and x 2 representing the domain R 2, and the third axis representing the codomain R. The graph of g is an upside down cone with vertex at the point bold u equals (u 1 comma u 2) and sides at angle of forty five degrees to the base plane. A point bold A equals (A 1 comma A 2) has been marked on the base plane and the image of bold A under the action of g is also indicated on the vertical axis. This is the point obtained by taking the square root of the sum of A 1 to u 1 all squared and A 2 to u 2 all squared.

Figure 1.3

A 3D graph showing the image set f of A. The impression is of a hollow cylinder, like a napkin ring, with the origin in the middle. Three mutually orthogonal axes are labelled as follows: x 1, a horizontal axis pointing towards us; x 2, a horizontal axis pointing right; and x 3, an axis pointing vertically up. Each x n is written x subscript n. The origin lies in the middle, hidden by the surface. The axis of the cylinder coincides with the vertical axis. The cylinder lies

symmetrically around the origin, extending a short distance above and below the horizontal plane.

Figure 2.1

A diagram of a triangle, illustrating the Triangle Inequality, M 3. The vertices are labelled lower-case A, b and c. The sides are labelled with the distances between the endpoints: d(A, b), d(b, c) and d(A, c). Inside the triangle, arrows parallel to the sides show the two routes from A to c, either directly or via b. The key point to notice is that the distance from A to c via b is longer than the distance from A to c directly along A c.

Solution to Exercise 2.1

A diagram of a triangle, illustrating that the distance d defined in the exercise cannot represent the length of the sides. The vertices are labelled x, y and z. The sides are labelled with numbers: x y is labelled 1, y z is labelled 2, and x z is labelled 4.

Figure 2.2

A graph showing a right-angled triangle to illustrate the taxi cab metric. Horizontal and vertical axes x 1 and x 2 are labelled, with the origin at the bottom left. Points A and b are marked, with A higher and left of b. Point b is labelled bold b = (b 1, b 2), and point A is labelled similarly. The coordinates are written with subscripts, for example A 1 as lower-case A subscript 1. Points A and b are joined with a solid sloping line. A right-angled triangle is formed by lines running right from A and up from b. The right angle is marked with a little box. The horizontal side is labelled modulus of b 1 minus A 1. The vertical side is labelled modulus of b 2 minus A 2. Thus the distance from A to b is measured by going across from A until above b, then down to b.

Figure 2.3

A 3D diagram illustrating the taxi cab metric, showing a few blocks of a city with streets laid out in a rectangular grid pattern, as in Manhattan, and the route taken by a taxi. An area of 3 by 3 squares, the blocks, is shown, plus an extra square adjoining the side nearest us. Point b is marked at the nearest corner, and point A at the far side. A route from A to b is shown by two thick lines, first going three blocks towards us until level with b, then turning left and going two blocks to b. Other possible routes through the grid are shown by dotted lines. All these routes would add up the same length. Buildings are drawn on three of the squares.

Figure 2.4

A graph showing the line x 1 + x 2 = 1. Horizontal and vertical axes x 1 and x 2 are labelled, with the origin at the bottom left. The points on the two axes one unit from the origin are labelled (1, 0) and (0, 1). These points are joined by a dotted line which is the line being illustrated and is labelled x 1 + x 2 = 1. This line forms a right-angled isosceles triangle with the two axes. A point on the line is labelled (x 1, 1 minus x 1). Each x n is written as x subscript n.

Figure 2.5

A graph showing the area where x 1 + x 2 lies between 0 and 1. It is similar to the previous figure except the triangle is shaded. Horizontal and vertical axes x 1 and x 2 are labelled, with the origin at the bottom left. The points on the two axes one unit from the origin are both labelled 1. These points are joined by a dotted line, forming a right-angled isosceles triangle as before. This triangle is shaded, and labelled 0 less than or equal to x 1 + x 2 less than 1. The dotted line indicates that this boundary is not part of the set. Each x n is written as x subscript n.

Figure 2.6

A graph showing the open ball centred on the origin with radius 1, with respect to the taxi cab metric e 1. The ball is not round but diamond shaped symmetrically around the origin. Horizontal and vertical axes x 1 and x 2 are labelled, with the origin at the centre. The four points on the axes one unit from the origin are labelled 1, 1, minus 1 and minus 1, and are joined by dotted lines to form the diamond shape, or a tilted square. The interior of the diamond, that is the ball, is shaded and labelled capital B subscript e subscript 1 (bold 0, 1). The dotted lines indicate that the boundary is not part of the ball.

Solution to Exercise 2.8

A graph showing the open ball with centre at the point (2, 3) and radius 2, with respect to the taxi cab metric e 1. As in the Figure 2.6, the ball is diamond shaped, this time with centre at (2, 3). Horizontal and vertical axes x 1 and x 2 are labelled, with the origin at the bottom left. The centre point, (2, 3), is marked. The four points lying two units straight up and down, and two units straight left and right from the centre form the corners of the diamond, or tilted square. These points are joined by dotted lines and the interior of the resulting diamond, that is the ball, is shaded. The ball is labelled capital B subscript e subscript 1 ((2, 3), 2). Dotted lines run across and down from the four corners and centre of the ball to the axes. The points where they meet the axes are marked: on the x 1 axis at 0, 2, 4, and on the x 2 axis at 1, 3, 5. The dotted lines forming the boundary of the ball indicate that the boundary is not part of the ball.

Figure 2.7

Part (a) of this figure illustrates the ball centred on the point bold A of radius r for the metric d when r is at most equal to one. If the point A has coordinates (A 1 comma A 2), then the ball consists of a line segment in the plane with left endpoint at (A 1 minus r comma A 2) and right endpoint at (A 1 plus r comma A 2). Neither endpoint is included in part of the ball. Part (b) of this figure illustrates the ball centred on the point bold A of radius r for the metric d when r is strictly larger than one. If the point A has coordinates (A 1 comma A 2), then the ball consists of the line segment in the plane with left endpoint at (A 1 minus r comma A 2) and right endpoint at (A 1 plus r comma A 2) together with the infinite vertical strip centred about the point bold A and of total width 2 r minus 2. The two vertical boundaries of the vertical strip are not included in the ball, except at the points where they

cross the horizontal line segment. The endpoints of the line segment are not included in the ball.

Figure 4.1

This figure illustrates the definition of continuity for a function f between two metric spaces X and Y. It consists of two shaded large blobs, one above the other. The top blob represents the metric space X and the bottom blob represents the metric space Y. There is an arrow indicating how f maps X into Y. In the space X there is a sequence of points labelled A 1, A 2, A 3 etc. that converge to a point labelled A. The image of this sequence of points, the sequence f(A 1), f(A 2), f(A 3) etc., is shown on the blob representing the space Y to converge to the point f(A) when f is continuous.

Chapter 15

Figure 1.1

This figure illustrates how the set X is embedded via the map f into the metric space W. It consists of two blobs with the first, representing X, placed above the other, representing W. Within W is a shaded region that represents the image of X under the map f. Points x and y in X are mapped by f to points f(x) and f(y) lying within W. The pullback of the metric d on W by f is the distance function on X that sets the distance between x and y in X to be the d distance between the image points f(x) and f(y) in W.

Figure 1.2

This figure shows the graph of the function f from R to R given by f of x equals arctan of x. The graph is strictly monotone increasing with horizontal asymptotes at pi over two and minus pi over two. The graph passes through the origin and is invariant under a rotation of 180 degrees. The tangent at the origin is a vertical line. As x tends to minus infinity, the function tends to minus pi over two; as x tends to plus infinity, the function tends to pi over two.

Figure 1.3

This figure shows the upper right quadrant of the plane, with a point P marked one unit up the vertical y-axis and points X and Y marked on the horizontal x-axis. The angle subtended by the line joining P to X and the vertical axis is alpha, and the angle subtended by the line joining P to Y and the vertical axis is called beta. The difference of the two angles, which is the angle subtended between the line segments P X and P Y, is called phi.

Figure 1.4

This figure shows the unit sphere with two points, bold x and bold y, marked on the surface. The distance between the two points bold x and bold y induced by the metric d 3 is the length of the chord that joins x and y. This chord is marked d superscript open-parenthesis three close-parenthesis subscript S (x, y) on the diagram. In contrast, the spherical distance

between bold x and bold y is the length of the shortest arc of the great circle that passes through both points.

Figure 1.5

S consists of three sides of a unit square: the left side, the base side and the right side.

Figure 1.6

This figure shows the different unit balls obtained for each of the metrics e 1, e 2 and e infinity.

The unit ball for e 1 consists of a filled diamond, that is, a rotated square, centred on the origin and with vertices at (minus 1, 0), (0, minus 1), (1, 0) and (0, 1). The boundary of the diamond is not part of the open ball and so is marked with a dashed line.

The unit ball for e 2 matches our usual idea of a ball and consists of a circular disc centred on the origin and with radius 1. The boundary of the disc is not part of the open ball and so is marked with a dashed line.

The unit ball for e infinity consists of a filled square centred on the origin, with side length 2 and with sides parallel to the x and y-axes. The boundary of the square is not part of the open ball and so is marked with a dashed line.

Figure 1.7

This figure shows the Hamming distance between pairs of points in A superscript 2. It consists of the four edges of a square together with its two diagonals. Each vertex of the square is one of the points of A superscript 2: The bottom left vertex is (0, 0); the bottom right vertex is (1, 0); the top right vertex is (1, 1); and the top left is (0, 1). The four bounding edges of the square are each marked with 1, which denotes the Hamming distance between the two vertices at the end of that edge. The diagonals of the square are each marked with 2, which is the Hamming distance between the points at the end of the diagonal.

Solution to Exercise 1.9

This figure shows the Hamming distances between pairs of points in A superscript 3. Rather than attempt to show these all on one unit cube, three cubes are used. Each cube consists of the twelve edges of a unit cube, with one vertex at the origin and all the edges parallel to one of the coordinate axes.

Points that are distance 1 apart are joined by black edges on the first cube and are all points directly joined by an edge of the cube. Points that are distance 2 apart are joined by blue (diagonal) edges lying in the faces of the second cube. Points that are distance 3 apart are joined by orange (diagonal) edges passing through the centre of the third cube. The second cube has twelve diagonal lines joining pairs of vertices that are distance 2 apart. The third cube has four diagonal lines passing through the centre of the cube joining vertices that are distance 3 apart.

Figure 2.1

This figure provides a visual representation of Cantor space by interpreting it as an infinite binary tree.

The figure looks a little like the canopy of a tree turned on its side. It starts at the left with a base node or vertex. From this vertex are drawn two edges with one going up and to the right and the other going down and to the right. The top edge is marked with a 0 and the bottom edge is marked with a 1. At the end of each edge is placed a further two edges also angled up and down and going to the right; these too are labelled: the upper ones with a 0 and the lower ones with a 1. This construction is then repeated ad infinitum. Thus, starting from the left of the diagram and moving right, we first see two edges, one labelled 0 and one labelled 1 and then four edges, labelled from top to bottom in the sequence 0 1 0 1, and then eight, sixteen etc.

A point in the Cantor space is represented by an infinite path starting at the far left of this figure and following appropriate branches to the right. Thus the point in Cantor space that starts 1, 0, 1 takes the lower edge at the far left, which is marked 1. Then at the junction it takes the upper edge, which is marked 0, and at the next junction it takes the lower edge, which is marked 1 and so on. This is marked on the figure by thickening the relevant branches.

The point 0, 0, 0, 0, dot dot dot is represented by the uppermost path, which is obtained by always following the top edge at each junction. The point 1, 1, 1, 1, dot dot dot is represented by the lowermost path, which is obtained by always following the bottom edge at each junction.

Figure 2.2

This figure illustrates three points in Cantor space, bold A, bold b and bold c. It is intended to help illustrate the proof of the Triangle Inequality in the case when bold A first differs from bold b at term m and bold A first differs from bold c at term n, and m is less than or equal to n.

Each point is represented as a pair of parentheses with filled circles, triangles or squares representing the values of each of the terms. Each point has a row to itself and the parentheses and shapes are aligned. There are two vertical dotted lines indicating the places at terms m and n in each of the points. The values of the terms of bold A are represented by filled circles. The values of the terms of bold b are indicated by filled circles up to the m minus 1 place, since bold A and bold b have the same terms up to this position. All subsequent terms of bold b are represented by blue triangles. Bold c lies beneath bold b and the terms up to the n minus 1 place are represented by filled circles, since bold A and bold c have the same terms up to this position. All subsequent terms of bold c are represented by blue squares.

Figure 2.3

This figure illustrates three points in Cantor space, bold A, bold b and bold c. It is intended to help illustrate the proof of the Triangle Inequality in the case when bold A first differs from bold b at term m and bold A first differs from bold c at the term n, and n is strictly less than m.

Each point is represented as a pair of parentheses with filled circles, triangles or squares representing the values of each of the terms. Each point has a row to itself and the parentheses and shapes are aligned. There are two vertical dotted lines indicating the places at terms n and m in each of the points. The values of the terms of bold A are represented by filled circles. The values of the terms of bold b are indicated by filled circles up to the m minus 1 place, since bold A and bold b have the same terms up to this position. All subsequent terms of bold b are represented by blue triangles. Bold c lies beneath bold b and the terms up to the n minus 1 place are represented by filled circles, since bold A and bold c have the same terms up to this position. All subsequent terms of bold c are represented by blue squares.

Figure 2.4

This figure illustrates the open ball of radius 2 to the minus n around a point bold A in the Cantor space. The figure consists of a grey image of the infinite binary tree that represents Cantor space as described in Figure 2.1. The ball consists of all branches of this tree that start A subscript 1, A subscript 2, dot dot dot, A subscript n. This is denoted on the figure by colouring red all those branches that start A subscript 1, A subscript 2, dot dot dot, A subscript n. This red figure looks like a tree on turned on its side, its trunk being the subbranch A subscript 1, A subscript 2, dot dot dot, A subscript n.

Figure 3.1

This figure shows how the balls for the different metrics e 1, e 2 and e infinity can be nested. Recall that the unit ball for e 1 is a regular diamond centred on the origin with its corners at (1, 0), (minus 1, 0), (0, 1) and (0, minus 1). The unit ball for e 2 is the usual disc centred on the origin and of radius 1. The unit ball for e infinity is a square of side length 2 that is centred on the origin and for which each side is parallel to one of the coordinate axes. The diagram shows that the unit ball for e 1 fits inside the unit ball for e 2, which in turn fits inside the unit ball for e infinity. Finally, the diagram shows that the ball for e 1 that is centred on the origin and has radius 2 contains all of the preceding balls. Note that the ball for e 1 that is centred on the origin and has radius 2 is a regular diamond centred on the origin with its corners at (2, 0), (minus 2, 0), (0, 2) and (0, minus 2). None of the balls include their boundary.

Figure 4.1

This figure shows the graphs of f 1, f 2, f 3 and indicates the graph of f n for large n. It consists of the right upper quadrant of the plane with both the x and y-axis coordinates running from zero to one. All the graphs of the functions contain the points (0, 0) and (1, 1).

Each graph in the sequence lies between the preceding graph and the x-axis. The graph of f 1 is just the straight line segment that joins the origin, (0, 0), to the point (1, 1). The graph of f 2 is a smoothly increasing curve with slope zero at the origin and slope two at the point (1, 1). The graph of f 3 has slope zero at the origin and slope three at the point (1, 1). The graph of the nth function has slope zero at the origin and slope n at the point (1, 1). Between these endpoints the slopes increase smoothly.

Figure 4.2

This figure shows the graph of the pointwise limit of the functions f n. The figure consists of the upper right quadrant of the plane with both x and y-axis running from zero to one. The graph of the pointwise limit f consists of the single point (1, 1) together with the line segment joining the origin, (0, 0), to the point (1, 0) that includes the origin but excludes the point (1, 0). Thus this line segment consists of all points of the form (x, 0) when zero is less than or equal to x which is strictly less than one.

Solution to Exercise 4.1

This figure shows the graphs of f 1, f 2 and f 3 on the right half plane with the x-axis running from zero to one. Each graph consists of two line segments, with the left line segment including both its endpoints and the right line segment only including its right endpoint. The left line segment is always a horizontal line starting at (0, 1) and the right line segment is horizontal and lies below the x-axis. For f 1, the left line segment joins (0, 1) to (a half, 1) and the right line segment joins (minus a half, a half) to (1, minus a half). For f 2, the left line segment joins (0, 1) to (a quarter, 1) and the right line segment joins (minus a quarter, a quarter) to (1, minus a quarter). For f 3, the left line segment joins (0, 1) to (an eighth, 1) and the right line segment joins (minus an eighth, an eighth) to (1, minus an eighth).

Solution to Exercise 4.2

This figure shows the graphs of f 1, f 2 and f 3 on the upper right quadrant of the plane with the x-axis running from zero to one. Each graph consists of line segments joined from end to end, starting at the origin and ending at the point (1, 0) on the x-axis. The first two segments make up the equal sides of an isosceles triangle with point upwards. For f 2 and f 3, a third segment is a horizontal line. For f 1, the peak of the triangle is at (a half, 2). For f 2, the peak is at (one quarter, 4). For f 3, the peak is at (one sixth, 6).

Figure 4.3

This figure illustrates how to interpret uniform convergence. It shows the graph of a continuous function f on the interval (0, 1) and the sleeve centred around it that is formed by shifting this graph vertically upwards and downwards by a distance epsilon. The graph of another function f n is sketched lying entirely inside this sleeve. The graph of f n lies entirely within the sleeve if, and only if, for each x in the interval (0, 1), the absolute value of f n of x minus f of x is at most epsilon.

Figure 5.1

The figure shows the graphs of two generic real-valued continuous functions from the interval (0, 1) and indicates where they have the largest difference: the absolute value (or modulus) of the size of this difference gives the max metric distance between the two functions.

Solution to Exercise 5.1(a)

This figure shows the graphs of f and g. It consists of the upper right quadrant of the plane with both x and y-axes running from zero to one. The graph of f is just a horizontal line segment joining (0, 1) to (1, 1). The graph of g is a diagonal line segment joining the origin to the point (1, 1). The maximum distance between f and g occurs when x is zero and is equal to one.

Solution to Exercise 5.1(b)

This figure shows the graphs of f and g. It consists of the right half plane with the x-axis running from zero to one and the y-axis running between minus one and one. The graph of f is just a horizontal line segment joining (0, a half) to (1, a half). The graph of g is a sine curve running through one complete cycle from x = 0 to x = 1. The maximum separation between f and g occurs when x is three quarters and is equal to three over two.

Solution to Exercise 5.1(c)

This figure consists of two separate diagrams. The first diagram shows the graphs of f and g. The diagram consists of the upper right quadrant with x and y-axes running from zero to one. The graph of f is the diagonal line segment that joins the origin to the point (1, 1). The graph of g is a smoothly increasing curve that lies below f and joins (0, 0) to (1, 1). The slope of g is horizontal at the origin and has slope larger than one at the point (1, 1).

The second diagram shows the graph of the modulus of h of x, where h is the function that is equal to f minus g. The diagram consists of the right half quadrant with x-axis running from zero to one. The graph of mod h looks like a smooth hill starting at height zero at the origin, rising to its highest point at x = a half and then descending again to meet the x-axis at x = 1. The slope at any point on the graph is always non negative and is horizontal at the highest point. This graph is then shown reflected about the x-axis using a dotted line labelled h of x.

Figure 5.3

This figure illustrates the effect of capital F subscript h on a real-valued continuous function f on the unit interval in the case when h is the function that maps x to x squared. The figure shows the graph h on the interval (0, 1) together with the graph of a generic function f. (In the particular case illustrated, the function f looks a little like a sine function.) In addition to the graphs of f and h, the graph of capital F subscript h of f is also indicated. Each y-coordinate on this graph is given by adding together the y-coordinates of the corresponding values on the graphs of f and h.

Figure 5.4

This figure shows three separate diagrams: the first shows the graph of f 1, the second the graph of f 2, and the third shows the graph of f 3. In each case the graphs lie within the upper right quadrant of the plane with x and y-axes ranging from 0 to one.

The graph of f 1 just consists of the line segment that joins the point (0, 1) to the point (1, 0). The graph of f 2 consists of a downwards sloping line segment joining (0, 1) to (a half, 0) together with a horizontal line segment joining (a half, 0) to (1, 0). The graph of f 3 consists of a downwards sloping line segment joining (0, one third) to (a third, 0) together with a horizontal line segment joining (a third, 0) to (1, 0).

Solution to Exercise 5.4

This figure shows the graphs of the functions f 1, f 2 and f 3. The graphs all lie within the upper right quadrant of the plane with x and y-axes ranging from zero to one. The graph of f 1 just consists of the line segment that joins the point (0, 0) to the point (a half, 1) together with the line segment that joins the point (a half, 1) to the point (1, 0). The graph of f 2 consists of three line segments: an upward sloping line segment joining (0, 0) to (a quarter, 1), a downwards sloping line segment joining (a quarter, 1) to (a half, 0) together with a horizontal line segment joining (a half, 0) to (1, 0). The graph of f 3 consists of three line segments: an upward sloping line segment joining (0, 0) to (a sixth, 1), a downwards sloping line segment joining (a sixth, 1) to (a third, 0) together with a horizontal line segment joining (a third, 0) to (1, 0).

Chapter 16

Figure 1.1

This figure illustrates the complement of a set A in an abstract space X. The space X is represented by a dark blue oval; the set A is represented by a smaller light blue oval lying wholly inside the dark blue oval. The complement of A, written A superscript C, is the dark blue region that is outside A but in X.

Solution to Exercise 1.5

This figure illustrates the argument of Exercise 1.5 and shows part of the real line surrounding the endpoint of the left closed interval lower-case A comma infinity. It shows the point x to the left of A and a point x subscript n to the right of A, and indicates that epsilon is chosen to be a small positive number that is less than A minus x.

Figure 1.2

This figure illustrates the argument of Worked Exercise 1.6 and shows the open ball with centre x and radius r together with the limit of the sequence lower-case A subscript n denoted by A and a term of the sequence A subscript n. Both A subscript n and A lie outside the open ball: the point A subscript n by assumption, and the point A because of the Reverse Triangle Inequality.

Figure 1.3

This figure illustrates the sets U and H of Worked Exercise 1.9. The set U is an open rectangle in the positive quadrant of the plane with left-hand corner at zero, zero. It is of width one and height three. The perimeter is dashed to indicate that it is not part of the set U. The set H is a closed rectangle in the positive quadrant of the plane with left-hand corner at zero, zero. It is of width one and height three. The perimeter is solid to indicate that it is included in the set.

Figure 1.4

This figure illustrates the sequence bold lower-case A subscript n used in part (a) of the solution to Worked Exercise 1.9. It shows the bottom left-hand corner of the set U with the origin and the point A subscript n marked. The point A subscript n has x-coordinate one over n plus one and y-coordinate one over n. It lies inside the set U.

Figure 1.5

This figure illustrates the sequence bold lower-case A subscript n that is used in the solution to Worked Exercise 1.10. It shows part of the open ball centred on x and of radius r together with the terms one and n of the sequence A subscript n. The first term is halfway along the horizontal line that joins the point bold x with coordinates x subscript one, x subscript two to the point with coordinates x subscript one plus r, x subscript two. The point bold A subscript n is indicated to lie on the same line segment between bold A subscript one and the point with coordinates x subscript one plus r, x subscript two.

Solution to Exercise 1.8

This is a sketch of the set A given in the exercise. It shows an infinite vertical strip of width 1 that includes the points (0, 0) and (1, 0), and also includes both of its boundary lines.

Figure 1.6

This figure shows the union of two sets A and B inside an overarching space X. The sets A and B are two partially overlapping grey shaded ellipses. The union consists of all points that lie in at least one of A and B and so consists of everything except those points that lie entirely outside both of A and B. These points are shaded light grey in the figure.

Figure 1.7

This figure shows that the points x and y will lie in different pieces of the level n approximation to the Cantor set if n is chosen so that three to the power of minus n is less than the modulus of x minus y. In the figure, x lies to the left, y lies to the right and indicative intervals from level n of the construction are drawn, one containing x and the other containing y. In between the two intervals is a gap and a point z is marked in this gap. z does not belong to the Cantor set.

Figure 2.1

This figure illustrates the proof of Theorem 2.7. It shows an open ball centred on a point lower-case A with radius r and which has a point x marked within it. Centred on x is drawn an open ball of radius s, where s has been chosen so that this ball lies entirely inside the original one. The point y is a point from the ball around x.

Figure 2.2

This figure illustrates the argument presented in the solution to Worked Exercise 2.10. It shows part of the open square S that is nearest the point lower-case A, which has coordinates (A subscript one, A subscript two), and illustrates that the open ball around A with radius r is completely inside the square S.(Recall that r is defined to be the smaller of 1 minus the modulus of A subscript one and 1 minus the modulus of A subscript two.)

Figure 2.3

This figure is associated with Worked Exercise 2.11 and illustrates those continuous functions that lie in the set A subscript c. The figure consists of a vertical segment that excludes its endpoints and is sited at the point c, is centred around c, and has total height two. All functions lying in A subscript c must cross this segment. The figure shows a function g that does lie in A subscript c and so g of c lies strictly between minus one and plus one, together with a function h that passes below the segment and so does not lie in A subscript c.

Figure 2.4

This figure is associated with Exercise 2.4 and illustrates the graphs of three continuous functions f, g and h from C zero, one to R, of which f and g lie in the set S and h does not. The functions f and g both take the value zero at zero and the function that does not, h, takes a positive value at zero.

Figure 2.5(a)

This figure illustrates De Morgan's first law for two sets A subscript one and A subscript two. The sets A subscript one and A subscript two are shown as two partially overlapping ellipses within an ambient space X. The complement of the union of A subscript one and A subscript two is shaded a lighter blue than either of the sets A subscript one and A subscript two and indicates the region under discussion in De Morgan's first law. It is the same as the intersection of the complement of A subscript one with the complement of A subscript two.

Figure 2.6(b)

This figure illustrates De Morgan's second law for two sets A subscript one and A subscript two. The sets A subscript one and A subscript two are shown as two partially overlapping ellipses within an ambient space X. Their intersection is shaded grey and every other region is shaded blue. The complement of the intersection of A subscript one and

A subscript two consists of everything that is not shaded grey. It is the same as the union of the complement of A subscript one with the complement of A subscript two.

Figure 2.6

This figure gives an example of a nested sequence of open intervals on the real line whose intersection, consisting of a single point zero, is closed. The open intervals are of the form minus one over n comma one over n for some natural number n. Each of them contains the point zero and their intersection contains nothing else.

Figure 2.7

This figure shows the sets U and J discussed in the main text. J is a closed line segment lying horizontally along the x-axis from minus two to plus two. U looks like a horizontal fat sausage symmetrically placed around J with rounded ends. It runs from minus three to plus three, has maximum height two and has a dashed boundary.

Solution to Exercise 2.10

This figure shows the shape W used in part (c) of Exercise 2.10. W is an open annulus centred at the origin with inner radius one and outer radius two. To show that the set W can be written as the union of open discs of radius one all centred on the circle around the origin of radius one, one such illustrative disc is sketched.

Figure 3.1

This figure attempts to illustrate that the closure of a set A is the smallest closed set containing it. It shows a collection of four sets with A the smallest of the sets and contained inside all of the other three sets. The second smallest also lies inside the other two larger sets and has a solid boundary. It is marked with a label showing that it represents the closure of A. The other two sets both contain A and the closure of A, and are labelled with the text ‘closed sets containing A’. They do not contain each other.

Figure 3.2

This figure illustrates how open balls around closure points always meet the set A. It shows a set A as a shaded ellipse inside the whole space X and also has two indicative points marked lower-case a and x. The point A is positioned inside A and the point X is positioned on the edge of A. Balls are drawn around both points and meet the set A.

Figure 3.3

This figure illustrates part of the solution to Worked Exercise 3.10. It shows that if there is a natural number n so that the point x lies in the open interval one over n plus one to one over n, then one can find an r greater than zero so that the open interval x minus r to x plus r also lies in the same interval.

Solution to Exercise 3.8

This figure shows an open square with centre at the point bold x with sides of length r that are parallel to the coordinate axes. There is a point in Q times Q marked within it. The whole square sits inside the open Euclidean ball of radius r that is centred on x, and so the point in Q times Q also lies in the ball.

Figure 4.1

This figure illustrates the relationship between the interior of a set and the set itself. It shows a set A that consists of several pieces, some of which are fat and have thickness and some of which are line like or dendritic in nature. The interior of the set A consists of the fat bits of A, excluding the boundary. Also on the figure is marked an open set U with dashed boundary that lies entirely inside one of the fat bits of A. This set U lies inside the interior of A.

Figure 5.1

This figure illustrates the meaning of being a boundary point for a set A. It shows a set A together with its complement, and a boundary point x. The boundary point x lies on the dividing line between A and the complement of A. An open ball centred on x is shown. It intersects both A and its complement.

Figure 5.2

This figure gives an abstract representation of the relationship between the following three disjoint sets. The interior of A, the interior of the complement of A, and the boundary of A. Together, their union forms the whole space X and the only other point of note is that A contains the interior of A and is disjoint from the interior of the complement of A.

Solution to Exercise 6.2

This figure illustrates the argument used in the solution of Exercise 6.2. It shows an open ball with centre x and radius r that meets the open set U, drawn as a large blob with dashed boundary. A point y lies in the intersection of the ball and the set U and it is possible to find a positive s so that the open ball centred on y and of radius s lies entirely within the intersection of U with the open ball with centre x and radius r. This ball must meet the open set V, since V is assumed dense in the metric space. A point z is chosen in this intersection.

Solution to Exercise 6.3

This figure shows a few points of the graph of the function defined in the solution to Exercise 6.3. The points 1 and 2 are mapped to 2 and 4, respectively. The points zero, minus 1 and minus 2 are mapped to the points 1, 3 and 5, respectively.

Figure descriptions for M303 Book E

Chapter 17

Figure 2.1

A set diagram showing different types of rings nested inside one another. Each type is represented by a rectangle, labelled with the type and an example. Six types are shown from the outermost, Rings, to the innermost, Fields. In order from the outermost the labels are as follows. First Rings with example Z subscript 6. Second Integral Domains with example Z[root minus 5]. Third Unique Factorisation Domains U F Dees with example Z[x]. Fourth Principal Ideal Domains P I Dees with example {half(a + b root 19 i) : a, b in Z). Fifth Euclidean Domains with example Q[x]. Sixth Fields with example Q.

Chapter 18

Nothing to describe.

Chapter 19

Figure 2.1

A geometric diagram showing two perpendicular lines and three intersecting circles with a list of four instructions for the construction. The line L is horizontal, with the point C above it. Instruction 1: Given a line L and a point C, draw a circle of centre C cutting L. Most of the circle lies above L. Instruction 2: Mark the intersection points with L as A and B. Point A lies to the left and B to the right. Instruction 3: Draw two circles with centres A and B and with the same radius as the first circle. These circles lie half above and half below the line L and intersect each other. Instruction 4: These circles will intersect at C and another point that we label as D. The line CD passes through C and is perpendicular to the line L. C D is a vertical line with C above L and D below L.

Figure 2.2

A symmetrical geometric diagram showing two perpendicular lines and two large intersecting circles and one small circle in their intersection, with a list of four instructions for the construction. The given line L is horizontal with the given point O in the middle of L. Instruction 1: Given a line L and a point O on L, draw a circle with centre O. This is the small circle. Instruction 2 (incorrectly given as the fourth instruction): Mark the intersection points with L as A and B. Point A lies to the left of O and B to the right. Instruction 3: Draw circles, centres A and B, with radius A B. These are the two large intersecting circles. The circle centre A passes through B. The circle centre B passes through A. Instruction 4 (incorrectly given as the second instruction): Mark the intersection points of these two circles as C and D. C D is a line through O and perpendicular to the line L. C D is vertical with C above L, and D the same distance below L.

Figure 2.3

A geometric diagram showing two lines intersecting at 60 degrees and two intersecting circles with a list of four instructions for the construction. The line L is horizontal. Instruction 1: Given a line L, mark two distinct points O and A on L. Point A is to the right of point O. Instruction 2: Draw a circle of centre O and radius O A. This circle passes through A and lies half above L and half below. Instruction 3: Draw a circle of centre A and radius O A. This circle passes through O and lies half above L and half below. The two circles are the same size and intersect. Instruction 4: These circles will intersect at a point that we label as B. Then the line through O and B is at angle 60 degrees to the line L. B lies above the line equidistant from O and A so that triangle O A B is equilateral.

Figure 2.4

A geometric diagram showing a circle with a right-angled triangle subtended on its horizontal diameter and the altitude of the triangle drawn. There is a list of five instructions for the construction. The heading of the instructions is: To draw a line whose length is the square root of the given length A B. Instruction 1: extend the line A B to C B so that A C = 1. Instruction 2: find the midpoint M of the line C B (see Exercise 2.3). Instruction 3: draw the circle of centre M and radius M B. Thus C B forms the diameter; C, A, M and B lie along the diameter from left to right. Instruction 4: draw the perpendicular to A B at A, and mark the intersection point with the circle as D. Thus D is on the circumference vertically above A. The lines C D and B D have been drawn forming a right-angled triangle on the diameter. Instruction 5: the length A D is the square root of the length A B, since A D divided by A C = A B divided by A D, and A D squared = A C times A B so that A D squared = A B as A C = 1.

Figure 2.5

A geometric diagram showing a complicated construction. The instructions are in the text following the figure. The point B is the top right corner of the given rectangle and is the centre of a circle. The rectangle is longer across than down. The corner below B is labelled A, and lies on the circle. The top left corner is labelled C and lies on a larger circle outside the first circle. The vertical line A B is extended to E on the large circle. The horizontal line C B extends right to D which lies on both circles and further right to the corner of the desired square. M is the midpoint of C D and the centre of the large circle. The chords C E and D E of the large circle are drawn. B is the bottom left corner of the desired square which extends to the right of its side B E.

Figure 2.6

A geometric diagram showing an obtuse-angled triangle with some comments. Detailed instructions follow the figure. The triangle is labelled A B C with base A B. The angle A is obtuse with side A C leaning up and to the left. Angles B and C are acute. Comment 1:point A 3 is the midpoint of A B. Comment 2: point A 6 is the midpoint of B C. These points are shown. A line joins A to its opposite midpoint, A 6. A line joins C to its opposite midpoint,

A 3. These two lines intersect at A 7 in the middle of the triangle. This point A 7 is the centroid. Another comment is: the construction lines for the midpoints have been omitted.

Figure 2.7

A symmetrical geometric diagram showing two intersecting circles with a list of five instructions for the construction. The two given points are labelled A and B. They are joined by a horizontal line with A to the left, and they are the centres of the two equal circles. Instruction 1: With centre A, draw a circle of radius A B. Instruction 2: With centre B, draw a circle of radius A B. Instruction 3: Label the two intersections of the circles as C and D. The line C D is vertical with C at the top. Instruction 4: The line C D is the perpendicular bisector to A B. Instruction 5: the midpoint M is the intersection of C D and A B. M is labelled. It lies in the centre of the diagram and is the desired point.

Figure 2.8

A geometric diagram showing two horizontal parallel lines cut by a vertical line with a list of four instructions. The given line L is the lower horizontal line. Point P lies above L where the upper line crosses the vertical line. Instruction 1: The line L and the point P are given. Instruction 2: Draw the perpendicular, M, from P to L. M is the vertical line. Instruction 3: Draw the perpendicular to M at P. This perpendicular is the upper horizontal line labelled L prime. Instruction 4: Then L prime is the line through P, parallel to L.

Solution to Exercise 2.2

A geometric diagram showing a complicated construction consisting of five circles and some lines. It has left–right symmetry. Detailed instructions are in the text following the figure. The horizontal lines L and L prime and the vertical line through point P are shown as in the previous figure. The vertical line is labelled M. Points A and B lie on L to the left and right of the vertical line. Three equal intersecting circles have centres at P, A and B. The circle centre P is in the middle and higher than the other two. Points on the vertical from top to bottom are labelled Y, P, X and D. Two larger circles have centres X and Y and radius X Y. The lower one with centre X encloses the three smaller circles. The small circle with centre P fits into the intersection of the large circles, touching both. The horizontal line L prime passes through the intersection points of the two large circles, with the point on the left labelled E. L prime is the desired line parallel to L, through point P.

Figure 2.9

A geometric diagram showing two circles and their common tangents. The smaller circle is on the left with centre O. The larger circle is on the right with centre A. There is a small gap between the circles. The line O A is horizontal and extends left of O, to a point X. Two tangents slope right from X, one above and one below the circles, touching both. The tangents form equal acute angles with the horizontal, above and below, so that the diagram has up–down symmetry.

Solution to Exercise 2.5(a)

A geometric diagram showing the construction of two intermediate points on the horizontal line O A. Detailed instructions are listed above the diagram. The initial line O A runs across the bottom with the origin O on the left. The line O to P 3 runs up and to the right from O at sixty degrees, so that O, P 3 and A form an acute-angled triangle. Two points, P 1 and P 2, are equally spaced between O and P 3. Two lines parallel to O P 3 through these points P 1 and P 2 meet the horizontal at the two desired points. Thus the large triangle is divided into stripes of equal depth parallel to A P 3. The desired points are labelled (a over 3, 0) and (2 a over 3, 0).

Solution to Exercise 2.5(b)

A geometric diagram showing the construction of n minus 1 intermediate points on the horizontal line O A. It is the same as the previous figure except that there are more points and more stripes. The line at sixty degrees runs from O to P n, with points P 1, P 2, P 3, and P n minus 1 in between. A dotted gap indicates that any number of points can lie between P 3 and P n minus 1. The horizontal runs from O to A as before. The desired intermediate points are labelled (a over n, 0), (2 a over n, 0), (3 a over n, 0) and ((n minus 1) a over n, 0).

Solution to Exercise 2.5(c)

A geometric diagram showing the construction of the tangent from a point P to a circle centre O. The circle is at the bottom left and the point P at the top right. Detailed instructions are listed above the diagram. The line O P is shown with midpoint M lying just outside the circle. A circle with centre M and radius O M is shown, a little larger than the first circle. It passes through O and the two circles intersect at two points. The upper point is labelled T. The desired tangent is the line P T.

Solution to Exercise 2.5(d)

A geometric diagram showing the construction of the common tangents to two circles. The diagram is based on the figure in the exercise but with additional lines and circles. Detailed instructions are listed above the diagram. O and A are the centres of the two original circles, with the smaller on the left with centre O. The right-angled triangle with base X O A is shown. Its apex is on the circle centre A, and its hypotenuse passes through the point on the circle centre O, directly above O. On the horizontal line X A from left to right are the points X, M, O, N and A. On the left a circle centre M has diameter X O. A large circle with centre N has diameter X A and encloses the two small circles on the left. The desired tangents pass through X and the intersection points of the pairs of circles. They slope symmetrically up and down from X and touch the circle centre A at points which are aligned vertically.

Figure 2.10

A geometric diagram showing a triangle with an extra line parallel to the right side, with a list of five instructions for the construction. The line O A D is horizontal across the bottom from left to right. Instruction 1: O A has length little a. The line O C B slopes up from O to the right at an acute angle. Instruction 2: O C has length 1 and O B has length little b. The line B D completes the large triangle O B D and the line C A completes the small triangle O C A. Instruction 3: D is constructed so that B D is parallel to C A. Instruction 4: Triangle O A C is similar to triangle O D B. Instruction 5: O A over O C = O D over O B and so O D = little a times little b.

Figure 2.11

A geometric diagram very like the previous figure. The only difference is that the labels A and D are interchanged, so that the base labels read O D A and the inclined side labels read O C B. The short line C D is parallel to the long side B A. Instruction 1: O A has length little a. Instruction 2: O C has length 1 and O B has length little b. Instruction 3: D is constructed so that C D is parallel to B A. Instruction 4: Triangle O D C is similar to triangle O A B. Instruction 5: O D over O C = O A over O B and so O D = little a over little b.

Solution to Exercise 2.11(a)

A geometric diagram showing the construction of a regular pentagon inscribed in a circle. Detailed instructions are listed above the diagram. A large circle centre O is shown, with vertical diameter A O B and horizontal diameter Y O X. (Incorrectly, Y is not labelled.) The point C lies on the vertical a little above O. The point D lies on the horizontal on the right, midway between O and X. The circle centre C radius C D lies inside the original circle, mainly in the top half. It cuts the vertical diameter a little below O at E. F is the point level with E on the original circle. It lies a little below X. The line F B is thus a chord of the original circle, on the bottom right. It is shown as a solid blue line and is the first side of the desired pentagon. The other sides are chords of equal length to F B and are shown as dotted blue lines. The pentagon has left–right symmetry with B at the bottom and the top side horizontal.

Solution to Exercise 2.11(b)

A geometric diagram showing another construction of a regular pentagon inscribed in a circle. Detailed instructions are listed above the diagram. A large circle centre O is shown, with vertical diameter A O B and horizontal diameter D O C. The point E lies on the vertical above O, midway between O and A. The line C E is the hypotenuse of a right-angled triangle with base O C. Two new smaller circles with centres E and C are shown. The circle centre E has diameter O A and touches the original circle at their highest point A. The point F lies on the hypotenuse C E about halfway along. The two new circles touch at F. The circle with centre C cuts the original circle at an upper point G and a lower point H. The line G H appears as a vertical chord of the original circle and is the first side of the desired pentagon. The other sides are chords of equal length to G H and are shown as dotted blue lines. The pentagon has up–down symmetry with the right side G H vertical.

Chapter 20

Figure 2.1

A graph of a cubic polynomial x cubed plus A times x plus b for A strictly greater than zero. The x and y axes, and origin O, are labelled. The graph consists of a continuous curve increasing from minus infinity on the left, towards plus infinity on the right. It rises steeply, cutting the x-axis to the left of the origin, then becomes much less steep, almost horizontal. It then cuts the positive y-axis and in the first quadrant it steepens again so that the overall shape is like a chair tilted back. The impression is that the curve would continue in this way to infinity. The curve cuts each axis only once.

Figure 2.2(a)

A graph of a cubic polynomial x cubed plus A times x plus b for A strictly less than zero. The x and y axes, and origin O, are labelled. The graph consists of a continuous curve. As in the previous figure it increases from minus infinity on the left, towards plus infinity on the right. The difference is that just above the x-axis there is a peak and then a trough before the curve rises steeply. It looks a little like a slice through a castle defence, with, going from left to right, a steeply rising hill followed by a moat followed by another steeply rising slope. The peak of the hill to the left and the trough of the ditch to the right are symmetrically placed around the y-axis. The particular instance of the curve shown here cuts each axis only once.

Figure 2.2(b)

A graph of a cubic polynomial x cubed plus A times x plus b for A strictly less than zero. The x and y axes, and origin O, are labelled. The graph consists of a continuous curve. As in the previous figure it increases from minus infinity on the left, towards plus infinity on the right, with a peak and a trough near the origin. The difference this time is that the curve has shifted downwards and cuts the x-axis three times, with (for this particular curve) the first two cuts occurring to the left of the y-axis and the third occurring to the right of the y-axis. The peak on the left and the trough on the right are equidistant from the y-axis.

Figure 2.3

A graph of a cubic polynomial x cubed plus A times x for A strictly less than zero. The x and y axes, and origin O, are labelled. The graph consists of a continuous curve. The curve is very similar to the previous figure, cutting the x-axis three times. The difference is that it passes through the origin instead of cutting the y-axis below the origin.

Figure 2.4(a)

A graph of a real elliptic curve from about x = minus 3 to about x = plus 4. The x and y axes, and origin O, are labelled. The graph consists of a continuous curve, symmetrical about the x-axis, looking rather like a narrow bell on its side, closed on the left and opening wider on the right. It crosses the x-axis once, at the peak of the bell. All of the curve lies to the right of this peak. The top half is now described in more detail; the bottom half is the mirror image. From the turning point of x on the x-axis at about x = minus 3, the curve first rises steeply but then flattens out at about x = minus 2 and continues at about the level y = 5. It crosses

the y-axis and then gradually steepens as the bell widens, reaching about y = 8 when it goes off the edge of the graph. The impression is that it would continue to rise towards infinity.

Figure 2.4(b)

A graph of a real elliptic curve from the y-axis to about x = plus 6. The x and y axes, and origin O, are labelled. The graph consists of a continuous curve, symmetrical about the x-axis. As in the previous figure there is a turning point of x on the x-axis, and the curve opens wider on the right. The main differences are that the curve is much straighter, looking almost like a V on its side, the turning point is much sharper, and the turning point is at x = 1. Thus all of the curve is to the right of x = 1. The curve gradually steepens as it widens. It goes off the right edge of the graph at about y = plus and minus 14. It crosses the x-axis once, at the vertex of the V.

Figure 2.5(a)

This graph appears the same as the V shaped curve in the previous figure, except that no values are marked on either of the x and y axes. The curve crosses the positive x-axis and lies entirely to the right of this point.

Figure 2.5(b)

A graph of a real elliptic curve. The x and y axes, and origin O, are labelled. The graph consists of a continuous curve, symmetrical about the x-axis. As in the previous figures, there is a turning point of x on the x-axis, and the curve opens wider on the right. The curve crosses the x-axis only at this turning point. The shape is more irregular, looking rather like a person's head, neck and sloping shoulders on its side, with the turning point at the top of the head. The top half is now described in more detail; the bottom half is the mirror image. From the turning point of x on the negative x-axis to the left of the origin, the curve rises steeply then flattens to a local maximum at just to the left of the origin. It then descends gently, crossing the y-axis a little above the origin. It reaches a local minimum just above the x-axis slightly to the right of the origin. This is the narrow neck. After the turning point it rises at about 45 degrees as the curve widens.

Figure 2.5(c)

A graph of a real elliptic curve. The x and y axes, and origin O, are labelled. The curve is symmetrical about the x-axis but this time the graph is not continuous but has two components. The difference from the previous figure is that the head is a closed curve separate from the shoulders, and the neck has gone. The curve cuts the x-axis at three points: at the top of the head, the chin, and the top of the shoulders. The top half is now described in more detail; the bottom half is the mirror image. From the turning point of x on the x-axis to the left of the origin, the curve rises steeply then flattens to a local maximum just to the left of the origin. It then descends, cutting the y-axis and reaching the x-axis again just to the right of the origin, so that this component is a closed loop. There is then a gap. The second component rises from further along the positive x-axis and rises steeply in an almost straight line to form a V shape.

Figure 2.6

A graph showing eleven isolated points. The y-axis is labelled from minus 5 to plus 5, and the positive x-axis is labelled from x = 1 to x = 10. The origin is not labelled. The graph is symmetrical about the x-axis. Five points lie in the upper half plane. These are from left to right (0, 1), (2, 3), (5, 4), (7, 5) and (9, 2). The mirror images of these five lie in the lower half plane. The point (10, 0) lies on the x-axis. Dashed lines run from each point to the axes to clarify their values.

Solution to Exercise 2.3

A graph showing eighteen isolated points. The y-axis is labelled from minus 6 to plus 6, and the positive x-axis is labelled from x = 1 to x = 12. The origin is not labelled. The graph is symmetrical about the x-axis. Nine points lie in the upper half plane. The nine points lie on just three horizontal lines, three at each of levels y = 1, y = 4 and y = 6. These are from left to right (1, 4), (2, 6), (3, 4), (4, 1), (5, 6), (6, 6), (9, 4), (10, 1) and (12, 1). The mirror images of these nine lie in the lower half plane. No points lie on either axis. Dashed lines run from each point to the axes to clarify their values.

Figure 2.7(a)

A graph showing the same elliptic curve as in Figure 2.5(c), the one with two components with the head separated from the shoulders. An extra straight line has been added with three points on the line labelled. These are the points where the line intersects the curve. The points P 1 and P 2 lie on the closed component of the curve near the top with P 1 to the left. The straight line P 1 to P 2 slices off a segment of the closed component. The straight line is extended to the right beyond P 2, and meets the second component at a point labelled pi (P 1, P 2). The straight line slopes up a little to the right. The line would not cut the curve again if extended indefinitely. The points P 1 and P 2 are written P subscript 1 and P subscript 2.

Figure 2.7(b)

A graph showing the same elliptic curve as in Figure 2.5(b), the one with a single component resembling a head, neck and shoulders. An extra straight line has been added in the upper half plane with three points on the line labelled. These are the points where the line intersects the curve. On the left ,the point P 2 lies on the head part near the top. On the right, the point P 1 lies on the shoulder part. The straight line slopes up a little to the right. The line cuts the curve in between P 1 and P 2 this time, on the head part a little to the left of the y-axis. This point is labelled pi (P 1, P 2). Thus the line slices off a segment of head, then runs above the neck before meeting the shoulder at P 1. The line would not cut the curve again if extended indefinitely. The points P 1 and P 2 are written P subscript 1 and P subscript 2.

Figure 2.8(a)

A graph showing the same elliptic curve as in earlier figures, the one with two components with the head separated from the shoulders. A different straight line has been added with two points on the line labelled. These are the points where the line intersects the curve. The

straight line slopes down to the right. On the left, the point P 0 lies on the head near its lowest point. The straight line is tangent to the curve at P 0. The line meets the shoulder component in the bottom right corner of the graph. This point is labelled pi prime (P 0). The line would not cut the curve again if extended indefinitely. The point P 0 is written P subscript 0.

Figure 2.8(b)

A graph showing the same elliptic curve as in earlier figures, the one with a single component resembling a head, neck and shoulders. A different straight line has been added with two points on the line labelled. These are the points where the line intersects the curve. The straight line slopes up to the right. On the right, the point P 0 lies on the neck a little above the x-axis. The straight line is tangent to the curve at P 0. The line meets the curve again at the bottom left of the graph. This point is labelled pi prime (P 0). The line would not cut the curve again if extended indefinitely. The point P 0 is written P subscript 0.

Figure 2.10(a)

A graph showing the same elliptic curve as in earlier figures, the one with a single component resembling a head, neck and shoulders. The straight line showing P 1 and P 2 with the third intersection point in between them has again been added. This line lies above the x-axis, slicing off a segment of head. As before, the third point is labelled pi (P 1, P 2). This third point lies on the head a little to the left of the y-axis. In this figure a new vertical dashed line is shown down from this third point. This vertical line meets the curve again below the x-axis at a point labelled P 1 plus P 2. It lies symmetrically below the x-axis from pi (P 1, P 2) on the opposite side of the head.

Figure 2.10(b)

A graph showing the same elliptic curve as in earlier figures, the one with two components with the head separated from the shoulders. The straight line showing P 1 and P 2 on the head with the third intersection point to the right of P 2 has again been added. As before the third point is labelled pi (P 1, P 2). This third point lies on the shoulder to the right of the y-axis. In this figure a new vertical dashed line is shown down from this third point. This vertical line meets the curve again below the x-axis at a point labelled P 1 plus P 2. It lies symmetrically below the x-axis from pi (P 1, P 2) on the opposite shoulder.

Figure 2.11(a)

A graph showing the same elliptic curve as in earlier figures, the one with two components with the head separated from the shoulders. The tangent from P 0 to pi prime P 0 has again been added. P 0 is on the lower half of the head, and the tangent slopes down to the right meeting the shoulder component in the lower right corner of the graph, at pi prime P 0. In this figure a new vertical line has been added up from this second intersection point. This vertical line meets the curve again at a point labelled 2 P 0. It lies symmetrically above pi prime P 0 on the opposite shoulder.

Figure 2.11(b)

A graph showing the same elliptic curve as in earlier figures, the one with a single component resembling a head, neck and shoulders. The tangent from P 0 to pi prime P 0 has again been added. P 0 is on the upper half of the neck, and the tangent slopes down to the left meeting the head in the lower left part of the graph at pi prime P 0. In this figure a new vertical line has been added up from this second intersection point. This vertical line meets the curve again at a point labelled 2 P 0. It lies symmetrically above pi prime P 0 on the opposite side of the head.

Figure descriptions for M303 History Reader

Section 1

Figure 1.1

This is a (possibly woodcut) portrait of Gerolamo Cardano with 'Hieronymus Cardanus Mediolanensis Medicinae Doctor Aetatis LXXI' written around the edge. The history of mathematics is sometimes (unfairly) described as being about dead white men with beards and this picture of a white middle-aged man with a beard does not do anything to refute this claim.

Figure 1.2

This is a portrait of Niels Henrik Abel. He does not have a beard and looks quite young.

Figure 1.3

This is a photographic portrait of Augustin-Louis Cauchy. He does not have a beard either.

Figure 1.4

This is a pen sketch of Évariste Galois. He is young and has no beard.

Figure 1.5

This is a photographic portrait of a bearded Camille Jordan.

Section 2

Figure 2.1

This figure shows the graph of the sum of the first one hundred terms of the Fourier series for the two pi periodic function given by f of x equals x over two for x between minus pi and pi. The graph is shown for x between zero and twelve point five and the jump discontinuities at odd multiples of pi result in vertical jumps in the graph. The sum of the first one hundred terms of the Fourier series appears to be a good approximation to the function except as x approaches one of the jump discontinuities, at which point there appear to be increasing oscillations that over- and undershoot the true target values. These oscillations increase in size as x approaches a discontinuity but appear to remain bounded in size.

Figure 2.2

This figure shows the graph of the Riemann function r of x on the open interval from minus two to plus two. R of x is a periodic function with period one and has a graph that consists of the line segment joining the point minus a half comma minus a half to the point plus a half comma plus a half. This is then repeated in each interval of length one. The drawing of the graph in the figure has vertical lines at minus one point five, minus point five, point five, and one point five. These are artefacts from the computer software used to draw the graph and are not part of the graph of the function.

Figure 2.3

This figure shows the graph on the interval minus one to one of the sum of the first one hundred and fifty terms in the series that defines the function f described in the text. This is a complicated graph with jump discontinuities at many points. It is reminiscent of the graph shown in Figure 2.2 but has many more sharp peaks.

Figure 2.4

This is the graph from minus pi to pi of the function given by sine pi times x plus sine of 2 times pi times x plus sine of 6 times pi times x plus sine of 24 times pi times x. The resulting graph is extremely complicated with some very large oscillations in it because each term in the sum can take value plus or minus one by choosing an appropriate value for x. Despite this, there are values of x for which the series is convergent. For example, the series converges when x is equal to zero.

Figure 2.5

This is a graph on the interval minus one to one of the sum of the first few terms of the Weierstrass function. The graph looks very much like a fuzzy, or fattened, sawtooth function with a lot of fine detail that is blurred out due to the resolution of the image. This is because the maximum magnitude of later terms in the series is small and so, although there is high-frequency oscillation in the later terms, they do not move far away from the sum of the earlier terms. As a result there is not a significant change in the appearance of the graph after adding only a few terms.

Figure 2.6

This is the graph of a false Weierstrass function on the interval minus one to one. This is a function that is of the same form as a Weierstrass function but for which the conditions on the coefficients that Weierstrass gave do not hold. It is visually rather different from the graph of a true Weierstrass function in that there appears to be a concentration effect going on and, rather than looking like a blurred sawtooth, the graph just looks very spiky with no blurring at all.

Figure 2.7

This is a photograph of a stone bust of Henry John Stephen Smith. He has a beard.

Figure 2.8

This is a figure showing two intersecting domains, T 1 and T 2. T 1 is a filled disc and T 2 is a filled square. The overlapping region is denoted T star. The part of the boundary of the disc that does not overlap with the square is called L zero. The part of the boundary of the square that overlaps with the disc is called L 1. The part of the boundary of the disc that overlaps with the square is called L 2. The part of the boundary of the square that does not overlap with the disc is called L 3.

Figure 2.9

This diagram has three subfigures. The first is of a disc with one radial spoke (aligned with the positive x axis) that starts at the boundary and finishes near the origin. The domain consists of the interior of the disc excluding those points lying in the spoke.

The second subfigure shows a disc with three radial spokes equally spaced around the boundary, one of which aligns with the positive x axis. All three spokes stop short of the origin. The domain consists of the interior of the disc excluding those points lying in the spokes. Thus to join a point that lies just near the boundary of the domain and just above the x axis to one that lies near the boundary but just below the x axis requires a path that is relatively long and passes near the origin.

The third and final subfigure shows a disc with radial slits alternating between the left- and right-hand halves of the boundary of the disc. The slits are interleaved so that in order to join a point near the top of the domain to a point near the bottom of the domain requires a long wriggly path that avoid each of the horizontal slits.

Figure 2.10

This is a photographic portrait of Maurice Fréchet. He wears round pebble glasses and has a moustache rather than a beard.

Figure 3.1

This is a portrait of Marin Mersenne. He appears to have a close-shaven beard.

Figure 3.2

This is a portrait of Pierre de Fermat. He looks rather like a Cavalier with long hair and no beard.

Figure 3.3

This is a portrait of Leonhard Euler. He has no beard and is wearing a rather floppy hat. His right eye appears to be closed and so this portrait was presumably painted after he lost the sight in his right eye.

Figure 3.4

This is a portrait of Joseph-Louis Lagrange in late middle age. He has no beard and a receding hairline. To make up for this, his hair is dramatically swept back.

Figure 3.5

This is an extraordinary caricature or cartoon showing Adrien-Marie Legendre and Joseph Fourier. Legendre's hair looks like three large white flames leaping upwards from his severe head. He has a stern expression and a thin, pointed face. In contrast, Fourier is shown with curly dark hair and a round face complete with double chin. He has a contented smile on his face. Around the two men are some mathematical sketches and notations.

Figure 3.6

This is a portrait of Joseph Fourier. He does not have a beard and appears to have been in early middle age when this portrait was painted. He has quite a round face, but the double chin shown in the previous caricature is missing!

Figure 3.7

This is a portrait of Carl Friedrich Gauss in his late middle age. He has no beard but large white mutton-chop sideburns. He is wearing a dark skullcap and is looking directly at the artist.

Figure 3.8

This is a portrait (possibly in pencil or charcoal) of Peter Gustav Lejeune Dirichlet. He is in his middle age and has a good-sized beard.

Figure 3.9

This is a photographic portrait of Bernard Riemann. He has pebble glasses and a full dark beard. He has a high forehead and probably has a comb-over. He is wearing a white shirt with a dark bowtie together with a double-breasted jacket.

Figure 3.10

This is a photographic portrait of Georg Cantor. He is a severe-looking bearded man who is nearly bald.

Figure 3.11

This is a photographic portrait of Felix Hausdorff and is unusual in that it is taken from the side. He has a magnificent goatee and is wearing a rather splendid dark tie under his jacket and waistcoat.

Good Academic Practice for M303

You should not copy solutions from the internet, or another student, for the following reasons.

1. It is unethical.
2. You will not learn anything by doing so. Learning maths is a little like learning to play a musical instrument in that you must *do* mathematics in order to learn it. Sending your tutor your own attempt at a question allows your tutor to spend time focusing completely on the way that you are doing mathematics and so give you individual one-to-one guidance and help. You are entitled to this one-to-one guidance and you will miss out on the opportunity to get it if you submit someone else's work.
3. At this level the material on the internet is often wrong (and it is fairly easy to spot a reproduced solution if it is incorrect).
4. If you get caught then the consequences can be very serious.

This does not mean that you must avoid looking at any mathematics on the internet, or working with someone else; but it does mean that you should not search the internet for answers to particular TMA questions. It also means that you should write your solutions to your TMA on your own, and 'in isolation', using only the module text (ideally using only the handbook as this is how you will do the exam).

When writing your solution, make sure that you fully understand how each line follows the last (if you are unsure then add a note to your tutor, explaining which bits you do not fully understand and asking for an explanation of those parts).

Formative Questions.

If you are stuck on a formative question then please *ask your tutor or ask on the forums for help.* Your tutor can help you with the formative questions and you are encouraged to discuss them with other students. However you should write up your solution on your own afterwards.

Summative Questions.

Your tutor cannot help you directly with these but it is still worth asking your tutor for help if you do not know where to start. If you are really stuck on all of the summative questions then contact your tutor and tell them.

There is additional university guidance on good academic practice in mathematics here.